NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

62-4-1

BOLT BERANEK AND NEWMAN INC

CONSULTING • DEVELOPMENT • RESEARCH

RADC-TDR-62-171

AN EVALUATION OF SPEECH COMPRESSION SYSTEMS

TECHNICAL DOCUMENTARY REPORT NO. RADC-TDR-62-171
1 March 1962

Rome Air Development Center
Griffiss Air Force Base
New York

Project No. 4519, Task No. 45350

(Prepared under Contract No. AF 30(602)-2235)

CAMBRIDGE, MASSACHUSETTS CHICAGO, ILLINOIS LOS ANGELES, CALIFORNIA

RADC-TDR-62-171

AN EVALUATION OF SPEECH COMPRESSION SYSTEMS

TECHNICAL DOCUMENTARY REPORT NO. RADC-TDR-62-171
1 March 1962

Rome Air Development Center
Griffiss Air Force Base
New York

Project No. 4519, Task No. 45350

(Prepared under Contract No. AF 30(602)-2235
by Kenneth N. Stevens, Michael H. L. Hecker,
and Karl D. Kryter of Bolt Beranek and Newman
Inc., Cambridge, Massachusetts)

TABLE OF CONTENTS

AN EVALUATION OF SPEECH COMPRESSION SYSTEMS

ABSTRACT

The results of PB word and nonsense syllable intelligibility tests, voice quality, talker identification, and continuous speech tests of selected speech compression systems are presented. The systems were: a "reference" low-pass (approximately 3000 cps) filter system, two channel vocoders, a semi-vocoder, a formant-tracking vocoder and a multiple narrow band filter system. The status of various speech compression techniques, current relevant research and recommendations for future research and development in this area are reported. Different speech compression techniques are classified according to their ability to provide a given level of speech intelligibility at different information rates.

It is judged that channel vocoders operating at about 2400 bits/sec and semi-vocoders at an estimated 9600 bits/sec provide adequate intelligibility and quality for most military communications; the quality of the semi-vocoder is superior to the channel vocoder. Formant-tracking vocoders utilize the lowest information rate (about 1000 bits/sec) of any of the bandwidth compression techniques. Formant-tracking vocoders require further improvement before they can be as satisfactory for general use.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The aims set forth for Contract USAF 30(602)-2235, "An Evaluation of Speech Compression Techniques," were:

(1) to determine the relative strength and weakness of presently available speech compression techniques,

(2) to evaluate these techniques as to possible future potential and expansion,

(3) to determine the best method for equipment development in the near future, and

(4) to determine the best areas for future intensive research effort.

It became apparent early in the work on the contract that these goals could not be met with confidence on the basis of existing information and published reports.[13,51*] For example, although steady progress has been made toward an understanding of the processes of human speech generation and perception over the past few years, the evaluation of speech compression procedures or techniques remains largely an empirical question. The capability of a given compression technique must be measured by an actual performance test and cannot be determined solely through an assessment of the technique in terms of some theory of speech. It also became obvious during the investigation that published test results of the performance

* References are listed in the Appendix, Section 5.1.

of individual speech compression systems must be interpreted with caution because of (a) the inherent unreliability of intelligibility test scores; (b) the influence on the test scores of extensive training of listeners for speech material processed by a particular system; (c) the somewhat unnatural mode of presentation of speech material in some performance tests; and (d) the frequent lack of a common reference system, by means of which one may compare the general capability of the talkers and the listening crew, and also study the effects of various test materials.

Accordingly, an important part of the present project was a testing program in which the performance of representative speech compression systems was measured by various types of listening tests. Following the testing program, an evaluation of the various speech compression techniques was made, both with respect to their suitability for immediate application and development and also with respect to their ultimate capabilities, possibly after several years of research and development. This evaluation was made partly on the basis of the results of the testing program, partly on the basis of published information on various speech compression systems, and partly on the basis of existing knowledge of the acoustics of speech and of the perception of speech.

The present report describes the various phases of the testing program in Section 2 and gives an interpretation of the test results in Section 3. A general evaluation of various speech compression techniques is presented in Section 4, together with recommendations concerning the present and future potential of the techniques. Several types of research studies that may contribute to the future development of new or improved speech compression techniques are also discussed in Section 4.

2. PERFORMANCE TESTS OF SELECTED SPEECH COMPRESSION SYSTEMS

2.1 List of Systems Tested

The following speech compression systems were tested in the course of the present study:*

(1) Reference (low-pass) system. This system consists of a Spencer-Kennedy Model 302 Electronic Filter, set for 1500 cps low-pass operation with a characteristic slope of -36 db/octave. The reference system is sometimes designated in this report as system R.

(2) Channel vocoder. Two systems were tested: Model HY-2, Philco Company, courtesy of the U. S. National Security Agency; and Model HC-135, Hughes Aircraft Company, Communications Division. Each vocoder has a 2400 bits/sec digital output, and an estimated 400 cps analog bandwidth. The philco vocoder is designated as system P; similarly, the Hughes vocoder is designated as system H.

(3) Semi-vocoder. General Dynamics Corporation, Stromberg-Carlson Division. This "base-band" vocoder has an estimated analog bandwidth of 900 cps and is designated as system S.

* A brief, functional description of these systems (except the Tasaroff-Daguet system) is given in Section 3 of this report. For reasons of security classification, the Tasaroff-Daguet system is described in a supplement of this report, Section 6. The results for the Tasaroff-Daguet system are presented along with results for the other systems in the body of the report, but the interpretation and evaluation of the data for that system are given in the supplement.

(4) Formant vocoder. Melpar, Inc. The formant-tracking vocoder tested produced a digital information stream of 1000 bits/sec; the bandwidth for analog operation is approximately 140 cps. The Melpar vocoder is designated as system M.

(5) Spectrum sampling (narrow-band) system. Bolt Beranek and Newman Inc. The analog bandwidth utilization is 800 cps. This "narrow-band" system is designated here as system N.

(6) Tasaroff-Daguet system. Courtesy of U. S. Army Signal Research and Development Agency. The estimated analog bandwidth is approximately 1000 cps. The Tasaroff-Daguet system is designated as system T.

These systems were selected because they represent examples of several different approaches to speech compression encompassing a range of bandwidths or digital transmission rates, and also because they happened to be available for testing.

Other methods for reducing the bandwidth required for transmitting speech have, of course, been developed or proposed. Some of these methods are discussed in Section 4 of this report. One technique of particular interest is the pattern-correspondence scheme, reported by C. P. Smith.[48,49] This system was not completely assembled at the time the present experiments were carried out, and hence could not be tested. It is understood, however, that tests comparing the pattern-correspondence scheme with more conventional vocoder methods will be carried out by C. P. Smith.

2.2 Summary of Tests

The speech compression systems listed above were subjected to several types of tests in the course of the present study. These tests were designed to measure:

(1) the intelligibility of phonetically balanced (PB) words,

(2) the intelligibility of nonsense syllables, with emphasis on the confusions made,

(3) the general quality of the processed signal,

(4) the accuracy with which listeners can recognize a given talker out of a small group of talkers, and

(5) the comprehension of continuous speech as a function of the degree of noise interference.

The reasons for selecting these particular types of tests will be discussed in the following sections. In general, however, the objectives were to devise a group of tests that could be related to other tests performed in the past, could provide some measure of the ability of talkers and listeners to communicate through the systems, and could give diagnostic information that would indicate any basic limitations in a particular system or would suggest modifications that may be made to improve system performance.

The general testing procedure was as follows: Various tests, designed to measure the factors indicated above, were recorded under laboratory conditions on magnetic tape. These test recordings were then played back through a number of speech compression systems, and the outputs of the devices were recorded on magnetic tape. The recordings of the outputs of the various systems were ultimately presented as tests to a crew of trained listeners under laboratory conditions.*

The prosecution of this test program was made possible through the cooperation of the various industries and governmental agencies responsible for the development and manufacture of the devices tested. All play-backs of the different speech tests through the various systems tested and all recordings of the outputs of these systems were made under the supervision of personnel responsible for the equipment, at the plant or laboratory where the equipment was developed.

2.3 Intelligibility Tests Using PB Word Lists

The phonetically balanced (PB) word test material consists of 1000 common monosyllabic words divided into twenty 50-item lists.[11] Each list contains the different types of speech sounds with a frequency of usage approximating that found in everyday American English. The score obtained on each list should, therefore, be a generally valid measure of the adequacy with which the communication system under test can handle everyday speech. Since the PB word tests consist of a large number of comparable lists, many conditions can be tested

* The listeners wore monaurally-fitted TDH-39 earphones made by the Telephonics Company; the earphones were calibrated on a 6 cc coupler and found to be flat within ± 5 db from 100 to 7000 cps.

in a single experiment without over-exposing the listeners to particular words. PB word tests have been so widely used that a standard has been prepared as a guide to their use by engineers.[1] These tests are typically scored in terms of the percentage of words recorded by the listeners that are totally correct.

Factors influencing intelligibility test scores. Scores for PB word tests are influenced by several factors that often make it difficult to compare the performance of speech communication systems that have been tested in different experiments. It has been demonstrated that the number of PB word lists actually used in testing a system has a significant effect upon the score obtained for that system.[37] For example, the score may be higher by as much as 20 percentage points if the listeners have been exposed to only four 50-word lists (or a total of 200 different items), than if all twenty 50-word lists (or 1000 different items) had been used. Because intelligibility tests reported in the literature indicate the use of different numbers of word lists, comparisons of the results of these experiments can be made only with considerable caution, if at all.

Another difficulty in the interpretation of intelligibility test results is that the scores tend to get progressively higher as the listeners gain more experience with a given system. This is particularly true for speech compression systems whose outputs have a peculiar sound quality. Under these conditions, with day-after-day training on a limited set of words and talkers, it is possible to obtain intelligibility scores that give an erroneous impression of the performance of the system for naive listeners.

There are two principal types of learning that seem to take place in a speech intelligibility test program. The first and most obvious is the learning of the voices of the particular talkers involved and of the speech material itself. In our experiments we attempted to overcome this type of learning by presenting to our listening crew for several weeks (three 3-hour sessions per week) special scramblings of our PB word and nonsense syllable tests. The reference system, a simple low-pass filter placed in an otherwise high-fidelity transmission link, was used for these tests, although a few tests were also given with a channel vocoder.

The second type of learning becomes evident as listeners become more and more familiar with the characteristics imposed upon different speech sounds by a particular system. Initially, the listening crew used in these tests was generally unfamiliar with the speech compression systems that we wished to evaluate. Although the amount of experience and listening afforded each system was approximately equal in the test evaluation program, the reference system, having also been used for training purposes, was presented more often. The scores obtained for the reference system probably benefited by an extra amount from both types of learning outlined above.

The effects of extensive experience with the reference system are illustrated in Fig. 2.3-1. The scores recorded during the initial training period and during the course of the experiment are shown for two signal-to-noise ratio conditions. From this figure we would estimate that the reference system scores for the experiment proper are probably 5 to 10 percentage points higher than they should be, relative to the scores obtained for the other speech compression systems, because of excessive exposure.

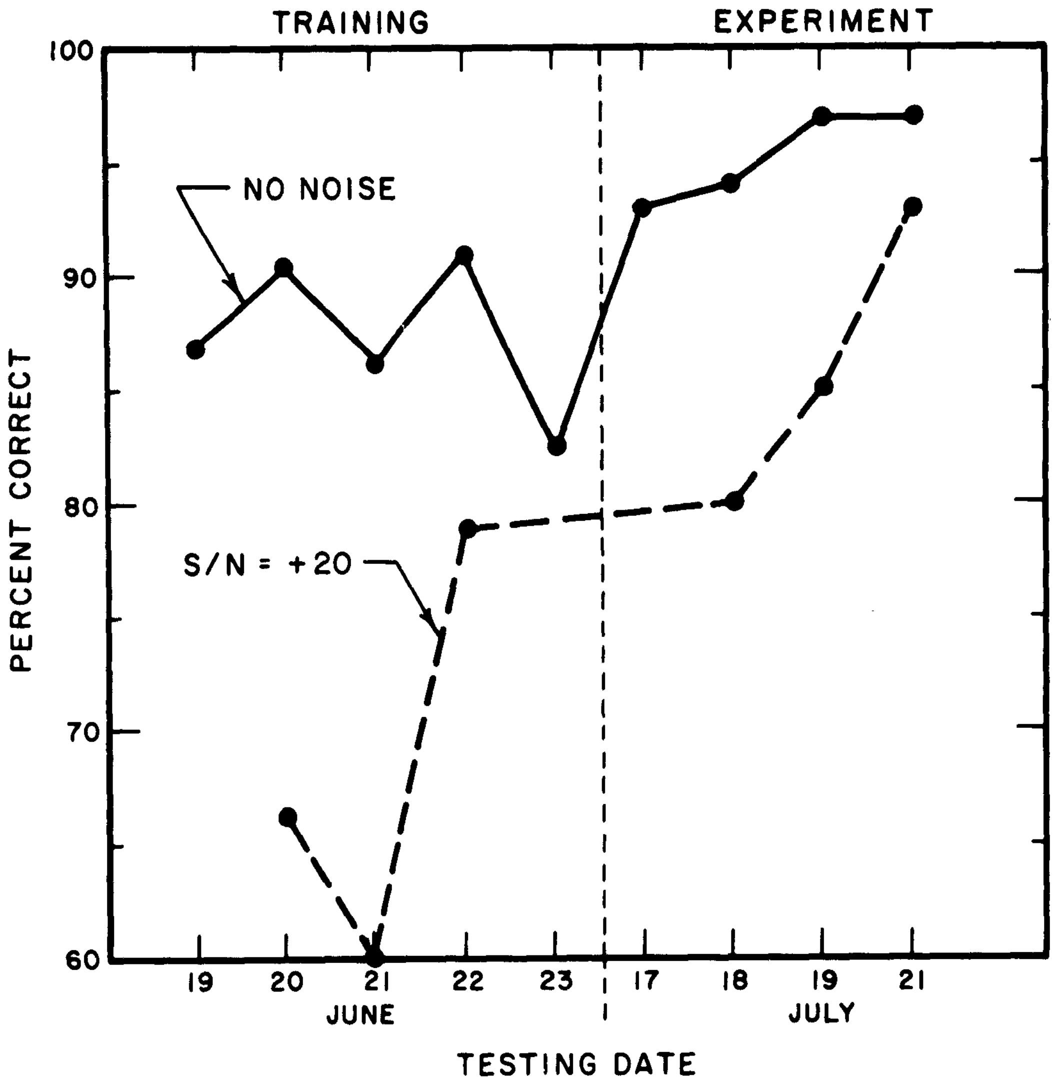


FIG. 2.3-1 AVERAGE SCORES OF PB LISTS WITH REFERENCE SYSTEMS
VOICE: MALE NO. 1

Design of the present test program. The recorded tests for each communication system to be evaluated were arranged to provide four sub-experiments. Each sub-experiment was designed so that the order of presentation of the tests obtained from the various talkers was randomly distributed among the systems. Through this procedure we feel that the effects of learning and order of presentation were approximately equally distributed among the various speech compression systems and that any one system was not favored by its position in the testing sequence.

The PB word tests that were prepared for determining the intelligibility of the speech compression systems may be conveniently divided into four groups:

Group	Talker	S/N Ratio	Microphone*	No. of Lists per System
I	Male 1	Optimal	Dynamic	4
	Male 2	Optimal	Dynamic	4
II	Female 1	Optimal	Dynamic	4
	Female 2	Optimal	Dynamic	4
III	Male 1	15 db	Dynamic	4
IV	Male 1	Optimal	Carbon	2
	Male 2	Optimal	Carbon	2

* The dynamic microphone used for making the test recordings was an Altec-Lansing Model 661A. The talker read the words in a soundproofed, semi-anechoic room, with the microphone positioned approximately 10 inches from his lips. The carbon microphone was that of a standard telephone handset, Western Electric Model 500. The talker held the handset in a normal position with the mouthpiece near his lips. All test recordings and system output recordings were made on either an Ampex Model 350 or Model 600 tape recorder operating at 7-1/2 i.p.s.

In groups I and II each talker recorded a different version of all twenty 50-word lists. A set of four of these recordings from each talker was then selected for processing by one of the seven speech compression systems. Other sets of four recordings were selected for each of the remaining systems, avoiding duplicate choices as much as possible. In group III the talker recorded another version of all twenty lists, this time against a constant background of filtered white noise. As before, a set of four recordings was reserved for each system. Complete versions of all twenty lists were not recorded in group IV, but a set of two recordings from each talker was chosen for processing by each of the four systems tested.

In addition to the tests described in the above four groups, two PB word lists were recorded by male talker No. 1 at Melpar, Inc., using their microphone facilities. While the recording was being made, the lists were processed "live" by the Melpar system; i.e., the microphone signal was recorded and simultaneously sent through the system. The recording of the microphone signal was then played back and processed by the Melpar system. Later, this same recording of the Melpar microphone signal was also processed by the Hughes system. On another occasion, male talker No. 1 read two different PB word lists live through the Philco system, using the microphone normally used with that system.

The Melpar and Philco systems were tested live in order to resolve a question which had been raised by some industries responsible for the development of speech compression devices regarding the comparability of recorded tests and live tests. The question may be considered to consist of three parts:

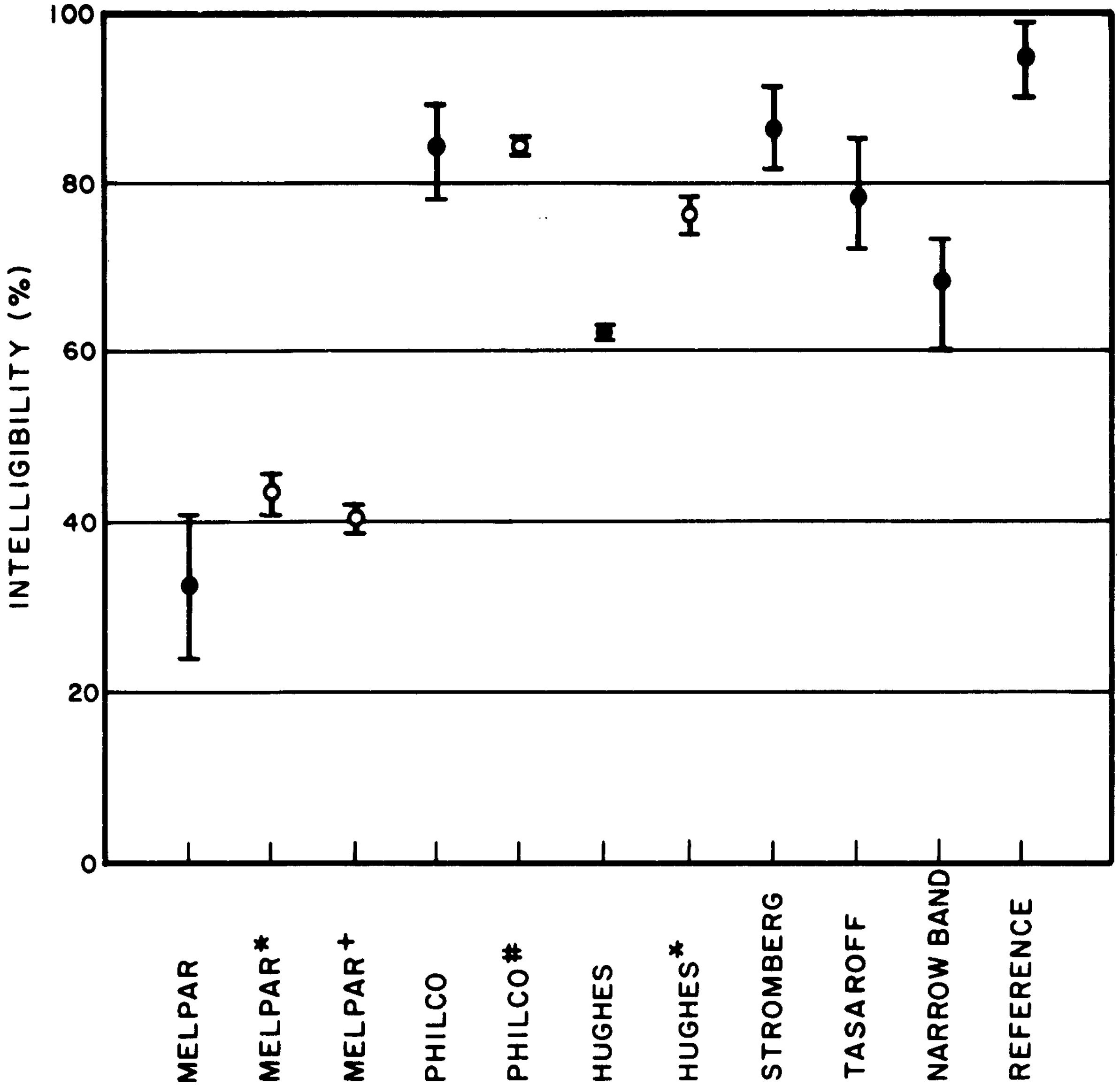
1. The inherent background noise and frequency distortion of the tape recording process could possibly reduce the performance of a sensitive system and thus degrade intelligibility scores;

2. The microphone used in making the tape recordings may have a frequency response that is significantly different from the response of the microphone for which a particular system is designed;

3. There may be a degradation in intelligibility because the operator of some systems normally holds the microphone within an inch or two from his lips, whereas the tape recordings were made with the microphone about 10 inches from the talker's lips.

Group I: Male talkers in quiet. The results of the PB word tests from Group I, averaged over the four sub-experiments and over a crew of eight listeners, are shown in Fig. 2.3-2 by the solid dots. The spread of the scores (averaged only over the listeners) within the sub-experiments are also indicated. This figure also presents the average scores obtained for the tests which were master recorded at Melpar, Inc., and for the tests recorded live with the Melpar and Philco systems. These scores are shown by the open circles.

Figure 2.3-2 shows that the intelligibility of the Melpar system improved through the use of a close-talking microphone, but also that the system was not adversely affected by using a tape recording as an input instead of a direct mirrophone. The Philco system is apparently less sensitive to the distance at which the microphone is used, and very similar scores are obtained whether the system is tested by means of our regular tape recordings or live with the close-talking Philco microphone. The overall signal-to-noise ratio was measured during



KEY

* MASTER TAPE MADE AT MELPAR, INC.
+ RECORDED "LIVE" AT MELPAR, INC.
# RECORDED "LIVE" AT PHILCO

FIG. 2.3-2 AVERAGE SCORES ON PB WORD TESTS
MALE TALKERS IN QUIET (GROUP I)

the recording of all master test tapes and was found to be in excess of 38 db. In contrast, the signal-to-noise ratio measured at the input of several systems that were operated live hardly approached and never exceeded this value. We therefore conclude that the recording process used does not have detrimental effects on the operation of the speech compression devices that were tested.

On the basis of the present test results it appears that the Melpar formant vocoder and the Hughes channel vocoder perform better when operated with a close-talking microphone instead of a microphone that is used at some distance from the speaker. A possible explanation for this is that some systems are more sensitive than others to the subtle spectrum differences between speech waves picked up close to the mouth (near-field condition) and those picked up at a point remote from the mouth (far-field condition). Also, it is reasonable to expect slight differences in the performance of a given system depending on the characteristics of the dynamic microphone being used.

Table 2.3-1 indicates which differences between the mean scores obtained for the seven rank-ordered systems are statistically significant. These data are based on an analysis of variance of test score distributions (see Table 5.2-1) and on the application of _t_ tests (see Table 5.2-2).

Table 2.3-1

Differences Between Scores Obtained for Group I: Male Talkers in Quiet (8 Subjects)

			Significant*
System	Rank	Average PB Score	Difference
Reference	1	95	
			Yes
Stromberg	2	86	
			No
Philco	3	85	
			Yes
Tasaroff-Daguet	4	79	
			Yes
Narrow Band	5	68	
			Yes
Hughes	6	61	
			Yes
Melpar	7	33	

(*Statistically significant at the $p \leq 0.01$ level of confidence.)

Group II: Female talkers. The results of the tests from Group II, again averaged over the sub-experiments and over the listeners, are given in Fig. 2.3-3. The Melpar and Tasaroff-Daguet systems were not tested for Group II. The spread of the scores over the sub-experiments is observed to be generally greater here than for the corresponding tests featuring male talkers.

Table 2.3-2 shows that the scores for the female talkers fall off more sharply for the channel vocoders and the semi-vocoder than for the reference and narrow-band systems. In the case of the channel vocoders this may be attributed to a difficulty in properly tracking the higher fundamental frequency of the female voices. In the case of the semi-vocoder the reduced intelligibility with female talkers may be explained in terms of the spectral location of the base-band. Fewer harmonics of the female voice are encompassed in this band, and hence it is more difficult to generate an excitation signal with a relatively uniform spectrum.

Table 2.3-2

Comparison of Average PB Word Scores for Male and Female Talkers

System	Male	Female	Difference
Reference	95	83	12
Narrow-Band	68	52	16
Philco	85	59	26
Hughes	61	34	27
Stromberg	86	56	30

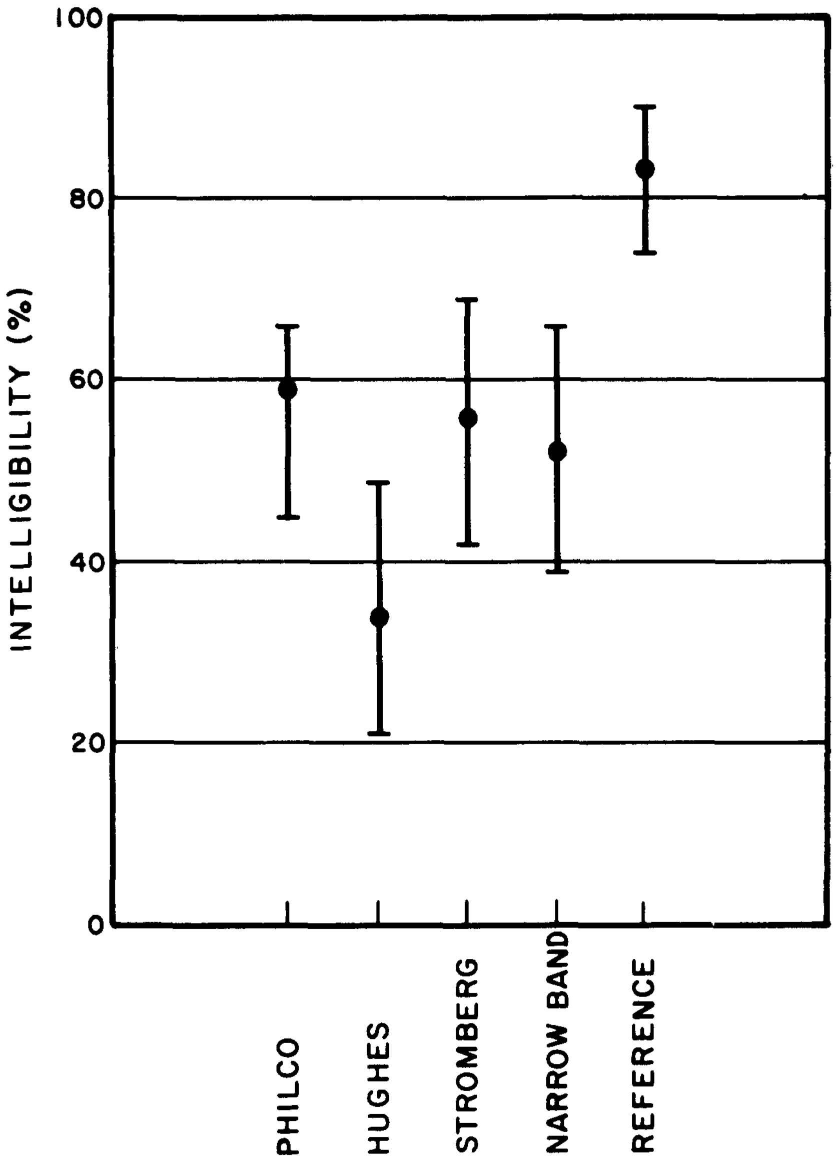


FIG. 2.3-3 AVERAGE SCORES ON PB WORD TESTS
FEMALE TALKERS IN QUIET (GROUP II)

Table 2.3-3 indicates which differences between the mean scores obtained for the five rank-ordered systems are statistically significant. The data are based on an analysis of variance of test score distributions (see Table 5.2-3) and on the application of t tests (see Table 5.2-4). The only two systems that do not score, on the average, significantly different from each other are the Stromberg semi-vocoder and the Philco channel vocoder.

Table 2.3-3

Differences Between Scores Obtained for Groups I and II: Male and Female Talkers in Quiet (8 Subjects)

System	Rank	Average PB Score	Significant* Difference
Reference	1	89	Yes
Philco	2	72	No
Stromberg	3	71	Yes
Narrow-Band	4	60	Yes
Hughes	5	47	

(*Statistically significant at the $p \leq 0.01$ level of confidence.)

From these statistical investigations it is evident that for the present PB word data a difference of 4 percentage points or more is significant. This is in general agreement with previous studies of PB word intelligibility tests, where it has been found that when averaged over 100 to 200 PB words (2 to 4 50-word lists) differences of 5 or more percentage points within a given experiment prove to be statistically significant.[11]

Group III: Male talkers in noise. A series of PB word tests were recorded with noise mixed in electrically at the input to the recorder. The noise was obtained from a white noise generator and was filtered to have a spectrum similar to the long-term average speech spectrum.[12] The signal-to-noise ratio was 15 db, as measured on a standard true RMS voltmeter set on "slow" meter action. The level of the speech was taken as the decibel average of the speech levels measured on each word in one PB word list. These recordings were made in order to demonstrate how the communication systems under test would perform if the talker had been in a moderate amount of ambient noise.

The results obtained with these test recordings are shown in Fig. 2.3-4. For comparison purposes, the average scores obtained for the quiet condition are also indicated in the figure.

The noise has a slight depressing effect upon the performance of the reference, narrow-band and Tasaroff-Daguet systems, and a drastic and harmful effect upon the performance of the Stromberg semi-vocoder, the Philco channel vocoder and the Melpar formant vocoder. On the other hand, the noise improved the performance of the Hughes channel vocoder. This finding was also borne out in the relative comprehension test, the results of which are given in Section 2.7. Comparative listening to the Hughes vocoder when operated in the quiet and in ambient noise produces the impression that the noise tends to stabilize the performance of the pitch extractor of the instrument. However, since the vocoder was operated in its digital mode, it is also possible that some misalignment in the digitizer circuits is responsible for this unexpected result.

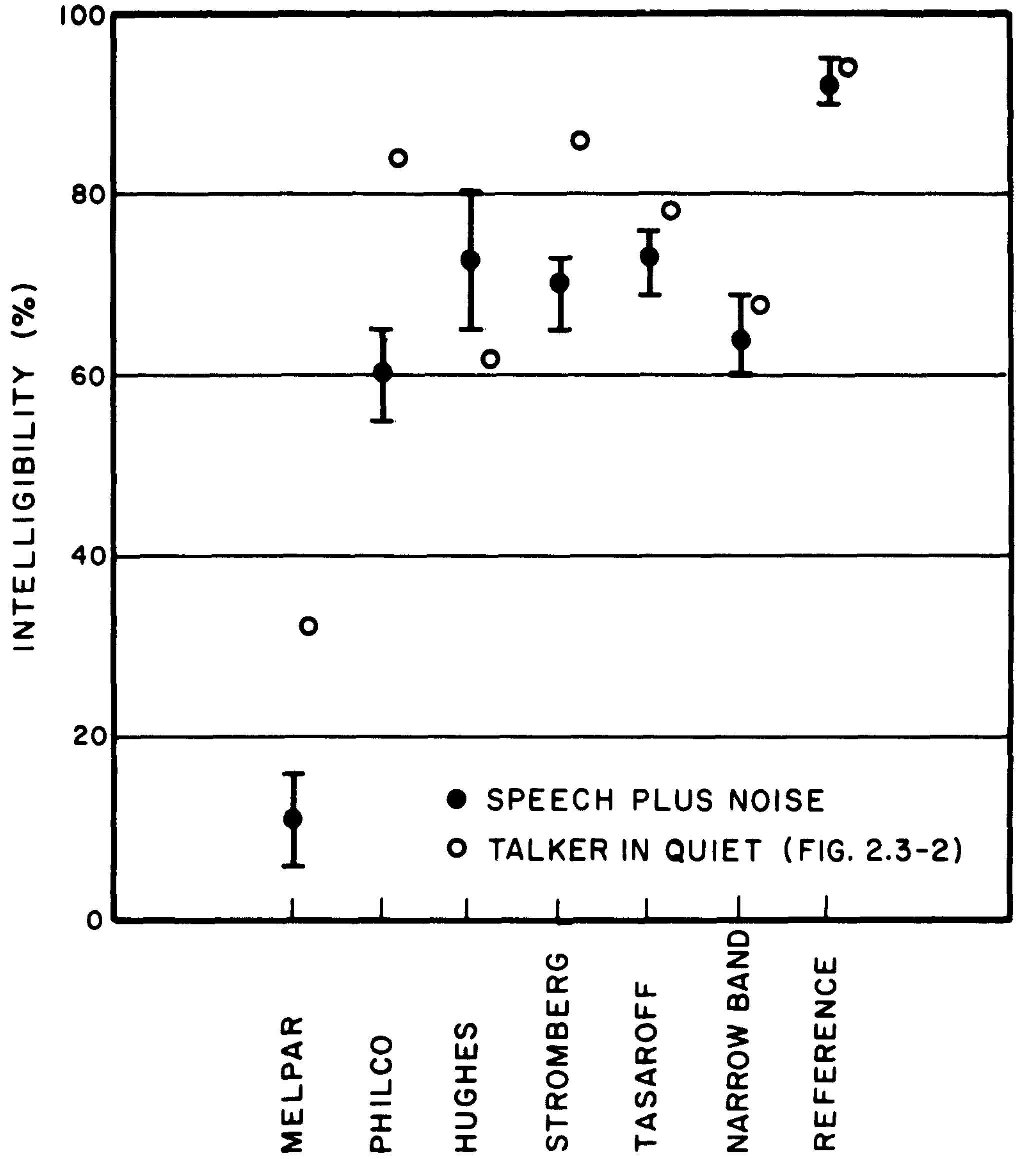


FIG. 2.3-4 AVERAGE SCORES ON PB WORD TESTS MALE TALKER, SPEECH PLUS NOISE (GROUP III)

Group IV: Male talkers in quiet, carbon microphone. Figure 2.3-5 shows the results obtained for male talkers when using a telephone-type carbon microphone. Also shown in the figure are average results obtained with the dynamic microphone. It is seen that the carbon microphone lowers the intelligibility scores by 4 to 10 percentage points. In view of the typically poorer frequency response of a carbon microphone in comparison to a dynamic microphone, such a reduction in scores is to be expected; that the scores are no lower than is indicated here is perhaps surprising.

2.4 Intelligibility Tests Using Nonsense Syllables

In the evaluation of speech compression systems we would often like to know in detail the performance of a system for various phonemes and classes of phonemes and for various distinctive features of the phonemes. This type of information may frequently help the experimenter to isolate the portion of a system that is responsible for a defect in performance and may lead to the design of suitable corrective measures. The information may also help to indicate any fundamental limitations in a particular speech compression technique. It may suggest, for example, that a particular technique is inherently incapable of making distinctions in the acoustic signal that are necessary cues for the identification of a particular distinctive feature of a phoneme.

In the type of intelligibility test that is generally used for diagnostic purposes the test material consists of nonsense syllables.[38] The nonsense material is usually monosyllabic, and consists of different vowels preceded and/or followed by a variety of consonants and consonant clusters. The number of consonants and consonant clusters that are used in American English is quite large (25-odd

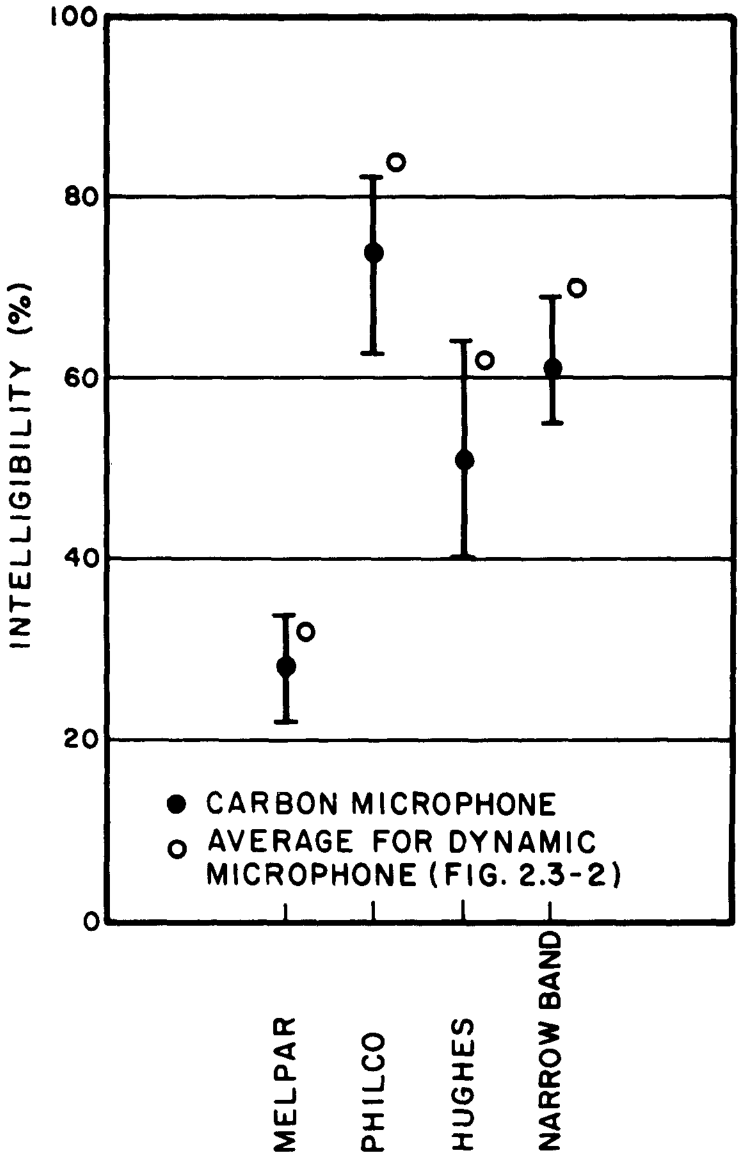


FIG. 2.3-3 AVERAGE SCORES ON PB WORD TESTS MALE TALKERS IN QUIET (GROUP IV)

consonants and an even greater number of clusters), and consequently a large number of syllables must be used if all situations are to be studied. In order to avoid an inordinate amount of testing time, some compromise is usually made and a shorter list of syllables is used. The total list of syllables is still, however, very long. Furthermore, highly trained listeners are required in this type of test, and the listeners must learn a long list of phonetic symbols or their equivalent.

As part of the present program of evaluating various speech compression systems, a group of new nonsense syllable tests has been developed. The approach that has been used represents one possible compromise to the problem of formulating suitable tests for diagnosing certain aspects of the performance of speech compression systems. Some of the tests are intended for the study of consonant intelligibility only; other tests evaluate vowel intelligibility. Each test in the group is quite short, and is designed to evaluate the performance of a speech communication link for only one or two consonant or vowel features. A large number of separate tests are therefore required to test a significant number of different vowel and consonant sounds; however, because each test is short and because each response must be selected from one of only a small set of possible responses, the listeners find the tests to be relatively easy to take, and the scores stabilize with very little training.

Consonant tests. A list of the consonants tested is shown in Table 2.4-1. Most, but not all, of the consonants of American English are included.

Table 2.4-1
Listing of Consonants Used in the Nonsense Syllable Tests
Arranged According to Place of Production (columns)
and Manner of Production (rows)

	A	B	C
	Bilabial Labio-dental	Post-dental	Alveolar Velar
1. Voiceless stops	p	t	k
2. Voiced stops	b	d	g
3. Voiceless fricatives	f θ (thin)	s	ʃ (shoe)
4. Voiced fricatives	v	z	ʒ (beige)
5. Nasals	m	n	ŋ (sing)
6. Glides	w (win)	j (yes)	
7. Liquids		ℓ	r

Some consonant clusters are tested in addition to the single consonants listed. The consonants in this table are arranged in a way that indicates roughly the manner of production according to rows, and the place of production according to columns. Thus Column A lists the bilabial and labio-dental consonants, all of which are produced with a vocal-tract constriction at the anterior end of the vocal tract. Column B lists the consonants that are produced with a constriction immediately behind the teeth, and Column C lists those that are produced with a constriction that is further back in the vocal tract, i.e., the alveolar and velar consonants.

The voiceless and voiced stop consonants are shown in rows 1 and 2 of the table, while rows 3 and 4 list the voiceless and voiced fricatives. The group /m n ŋ/ in row 5 are nasals, /w j/ in row 6 are called glides, and /ℓ r/ in row 7 are called liquids.

The list of utterances that constitute each of the consonant tests contains several versions of each of a relatively small number of consonants (four to eight) occurring in initial and/or in final positions in syllables. Each list is further simplified by including only two possible syllabic nuclei or vowels. One of the vowels is always a long vowel and the other is short; one is a back vowel and the other is a front vowel. Four lists of syllables, differing only in the pair of vowels used, are assembled for any one group of consonants. The vowels are always selected from the set /i ɪ ɛ æ ɑ ʌ ʊ u/, and four different pairs of these vowels are selected to assemble test lists for each consonant group. Within any one list a given consonant generally appears once in initial position and once in final position with any one vowel.

In all, nine groups of consonants were tested, and hence the total number of syllable lists was 36. The nine groups of consonants are listed in Table 2.4-2. In this table, tests 1a, 1b, 1c and 1d, for example, are the four tests that examine the six consonants /p t k b d g/; the vowel pairs for this set of tests were, respectively, /i ʌ/, /æ ʊ/, /ɪ ɑ/ and /ɛ u/. The final column in the table indicates for each row the particular features that are tested by examining responses to the group of consonants in that row.

The test syllables in all cases except the final clusters are preceded by an unstressed carrier syllable /h ə/. Thus, a typical item in test 1a would be /h ə ˈp ʌ d/. In the case of the tests of final consonant clusters, monosyllabic utterances were used, preceded by the consonant /h/. Thus, a typical item in the group of tests 8a-8d would be /h ɛ n t/.

Vowel tests. Eight different vowels were used in these tests -- the vowels /i ɪ ɛ æ ɑ ʌ ʊ u/. The tests were divided into four groups of two tests each, for a total of eight tests; four vowels were included in each of the groups. A listing of the vowels for each group is given in Table 2.4-3.

Table 2.4-2

Groups of Consonants Contained in Individual Nonsense Syllable Tests

Tests	Consonants	Features Studied
1a-1d	p t k b d g	Voiced-voiceless for stop consonants; place for stop consonants
2a-2d	p t k f s ʃ	Interrupted-continuant for voiceless consonants; place for voiceless consonants
3a-3d	b d g v z ʒ	Interrupted-continuant for voiced consonants; place for voiced consonants
4a-4d	f s ʃ v z ʒ	Voiced-voiceless for fricative consonants; place for fricative consonants
5a-5d	b d v z m n	Manner for voiced consonants; place for voiced consonants
6a-6d	f θ s ʃ	Place for voiceless fricatives
72-7d	w j m n r ℓ (initial) m n ŋ r ℓ (final)	Manner for voiced consonants; place for voiced consonants
8a-8d	s t n ℓ r rt st ℓt nt	Final clusters
9a-9d	s sp st sw sℓ sm sn str	Initial clusters

Table 2.4-3
Vowels Used in Each of the Four Groups of Vowel Tests

Test	Vowels	Description of Vowel Series
10a, 10b	i ɪ ɛ æ	Front vowels
11a, 11b	ɑ ʌ ʊ u	Back vowels
12a, 12b	i æ ɑ u	Long vowels
13a, 13b	ɪ ɛ ʌ ʊ	Short vowels

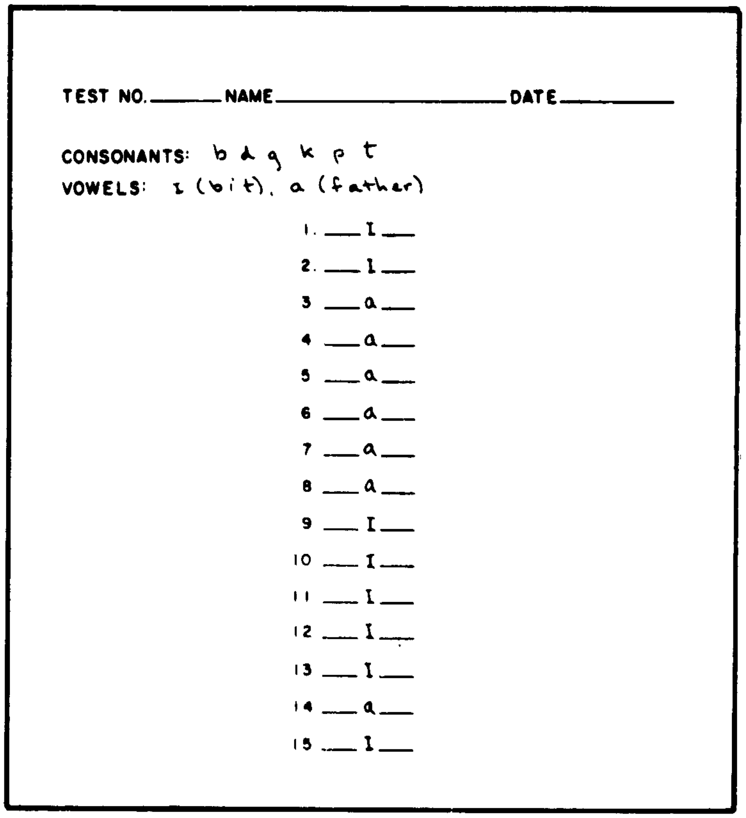
Phonetic symbols:
i (beat)
ɪ (bit)
ɛ (bet)
æ (bat)
ɑ (father)
ʌ (but)
ʊ (foot)
u (boot)

The structure of each test item was similar to that of the items for most of the consonant tests, i.e., each utterance consisted of the unstressed carrier syllable /h ə/ followed by a stressed consonant-vowel-consonant syllable. The initial and final consonants were identical for a given test item. For the "a" series of tests (i.e., 10a, 11a, 12a and 13a) the consonant environments were selected from the voiceless consonants /p t f s/; the voiced consonants /b d v z/ formed the environments for the "b" series of tests. Thus a typical utterance in test 10a, for example, was /h əˈ f ɪ f/. In each test, each vowel occurred once in each consonant environment, so that there were 16 test syllables in all.

Preparation and administration of tests. As noted above, the number of test items in each test is determined by the number of consonants and vowels that are included in the test. Thus, for example, tests la-ld, which each include 6 consonants in 2 vowel environments, have a total of 12 test items each; each vowel test includes 4 vowels in 4 consonant environments, giving a total of 16 test items. In the preparation of the test lists, three additional nonsense syllables were added to each list of test items in order to destroy the apparent symmetry of the tests. These "dummy" items were distributed randomly throughout each list and were not included in the scores for the test.

Each of the 44 tests was recorded by two talkers, different random orders being used for each talker. The talkers were trained in generating these types of utterances, and were instructed to read the items with constant voice effort (rather than using a VU-meter to monitor the level) and, as far as possible, with the same inflection for each item. All recordings were later monitored by both talkers, and, in the case of utterances that were judged to be unacceptable, new recordings were made. These recordings were processed by the various speech compression systems that were being evaluated, except for the Tasaroff-Daguet system. A more restricted set of tests was processed by the Tasaroff-Daguet system, as discussed below.

In the administration of the tests, the listeners were provided with answer sheets of the type shown on Figs. 2.4-1 (typical consonant test) and 2.4-2 (typical vowel test). For a given consonant test, the possible consonant responses are indicated at the top of the sheet. For each test item, the vowel is given, and blanks indicate where the consonant responses are to be entered. In the case of the vowel tests, the possible vowel responses are indicated at the top of the sheet and the consonant environment is given for each syllable.

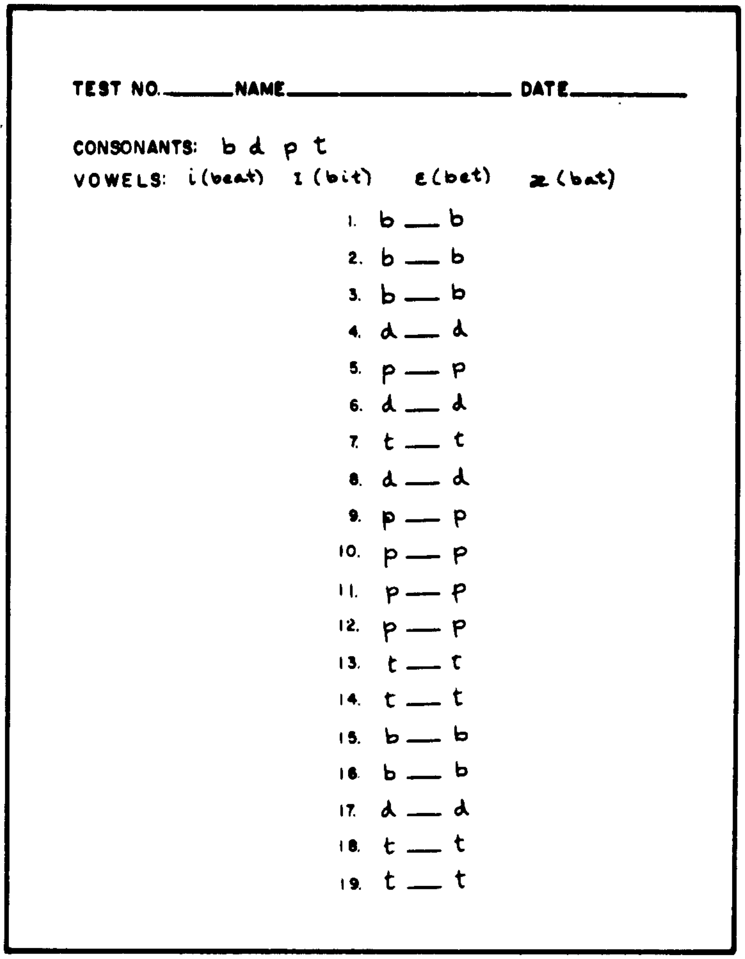


TEST NO.______ NAME______________________ DATE__________

CONSONANTS: b d g k p t
VOWELS: ɪ (bit), a (father)

1. ___ I ___
2. ___ I ___
3 ___ a ___
4 ___ a ___
5 ___ a ___
6 ___ a ___
7 ___ a ___
8 ___ a ___
9 ___ I ___
10 ___ I ___
11 ___ I ___
12 ___ I ___
13 ___ I ___
14 ___ a ___
15 ___ I ___

FIG. 2.4-1 TYPICAL ANSWER SHEET USED IN NONSENSE SYLLABLE TESTS FOR CONSONANTS



TEST NO.______NAME____________________ DATE__________

CONSONANTS: b d p t
VOWELS: i (beat) ɪ (bit) ε (bet) æ (bat)

1. b __ b
2. b __ b
3. b __ b
4. d __ d
5. p __ p
6. d __ d
7. t __ t
8. d __ d
9. p __ p
10. p __ p
11. p __ p
12. p __ p
13. t __ t
14. t __ t
15. b __ b
16. b __ b
17. d __ d
18. t __ t
19. t __ t

FIG. 2.4-2 TYPICAL ANSWER SHEET USED IN NONSENSE SYLLABLE TESTS FOR VOWELS

The listening crew was given several training sessions for nonsense syllables before the test material from the various speech compression systems was administered. The first few training sessions indicated some improvement in the overall performance of the subjects, but the scores reached relatively stable values before the tests proper were begun. For the vowel tests, phonetic symbols were used to indicate responses, but only about two hours of training were required before these symbols were learned adequately. The problem of learning the phonetic symbols is, of course, minimized when only four possible responses are required in a given test.

The same crew of eight listeners was used for both the PB word and the nonsense syllable tests. Suitable precautions were taken to balance out any learning that might have occurred throughout the test series, as discussed in Section 2.3 for the PB word tests.

Nonsense syllable tests used with the Tasaroff-Daguet system. In the overall testing program, it was necessary to process the speech recordings by the Tasaroff-Daguet system several months before the other systems were tested. At the time the Tasaroff-Daguet tests were made, the complete set of nonsense syllable tests described above was not available, and it was therefore necessary to use a more restricted group of recorded tests, since these were all that were available. These tests had been recorded by the same talkers who recorded the subsequent more extensive series of tests. The restricted group of tests were designed to evaluate consonant intelligibility only; no vowel tests were available for processing by the Tasaroff-Daguet system. These materials included only one or two tests for each consonant group (rather than four, as in the extensive tests), and hence not all vowel environments were included. Also, the particular consonants included in two of the tests were slightly different from those in Table 2.4-2.

Results. The results of the nonsense syllable tests are presented in Tables 2.4-4 through 2.4-10 as a series of confusion matrices. These matrices in which the entries represent percentages, summarize results pooled for the two talkers and for all listeners, and combined for initial and final consonant positions where applicable. Twenty-four confusion matrices are given for each system (except the Tasaroff-Daguet system). The test data that were pooled to obtain these confusion matrices are listed in Table 2.4-11.

The number of individual responses on which an entry in a confusion matrix is based varies from 144 for the vowel tests 10-13 to over 1500 for the voiced-voiceless distinction.

The data from the confusion matrices in Tables 2.4-4 through 2.4-10 are still further collapsed in Table 2.4-12, which lists the average percentage error for each confusion. Each entry in this table, when divided by 100, can be interpreted as the probability that a stimulus is categorized incorrectly, i.e., the probability that a response occurs as any off-diagonal entry in the relevant confusion matrix. The results for the Tasaroff-Daguet system are derived from a restricted series of tests, and cannot be compared directly with data for the other systems.

Table 2.4-13 shows the result of averaging data from groups of features. Some of these averaged data are plotted in the various portions of Fig. 2.4-3. The PB word scores for male voices in quiet, previously shown in Fig. 2.3-2, are replotted for comparison with the various nonsense syllable scores.

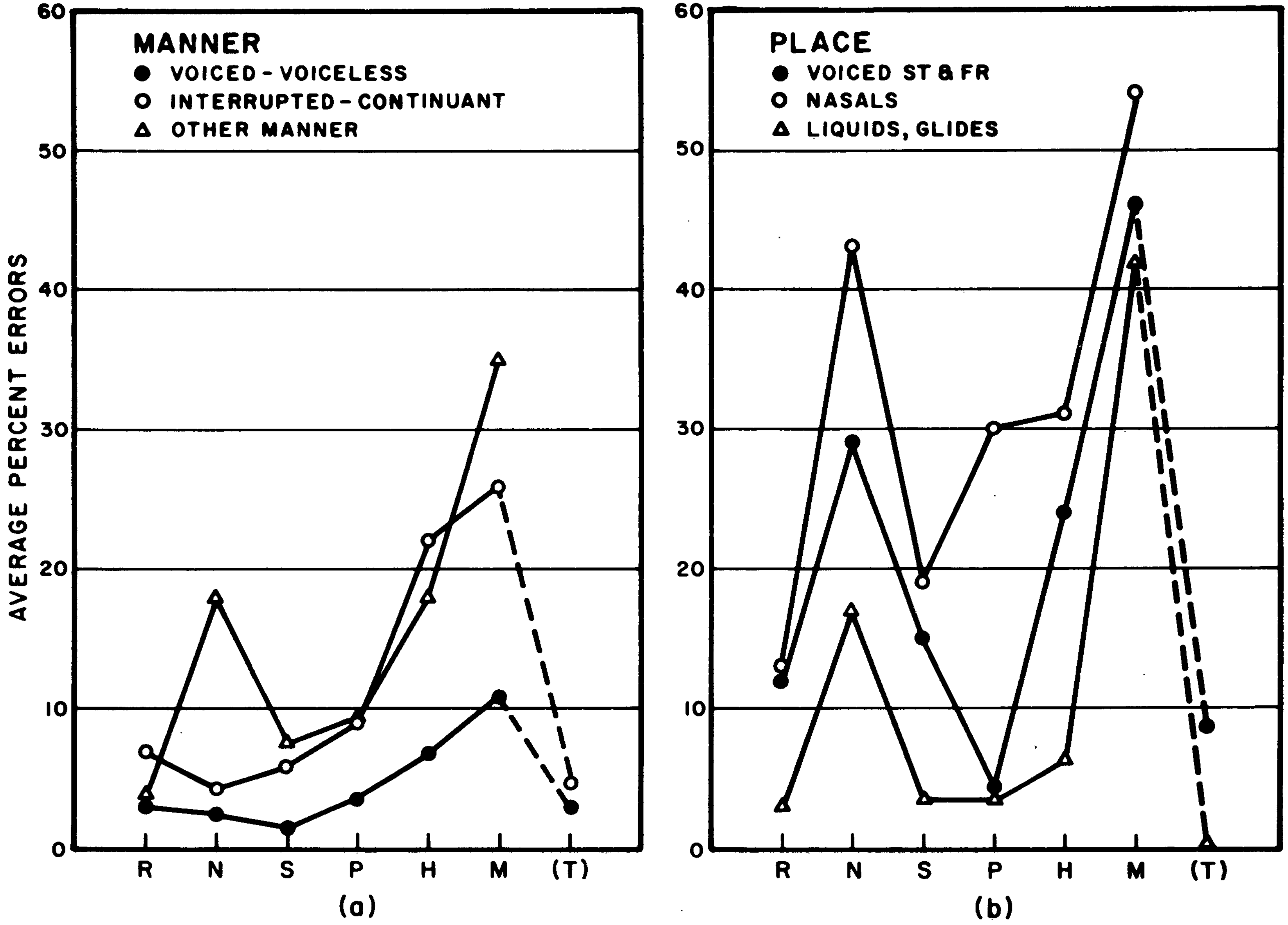


FIG. 2.4-3 AVERAGE PERCENT ERRORS FOR DIFFERENT CONSONANT FEATURES, AS DERIVED FROM RESULTS OF NONSENSE SYLLABLE TESTS. SEE TABLE 2.4-13

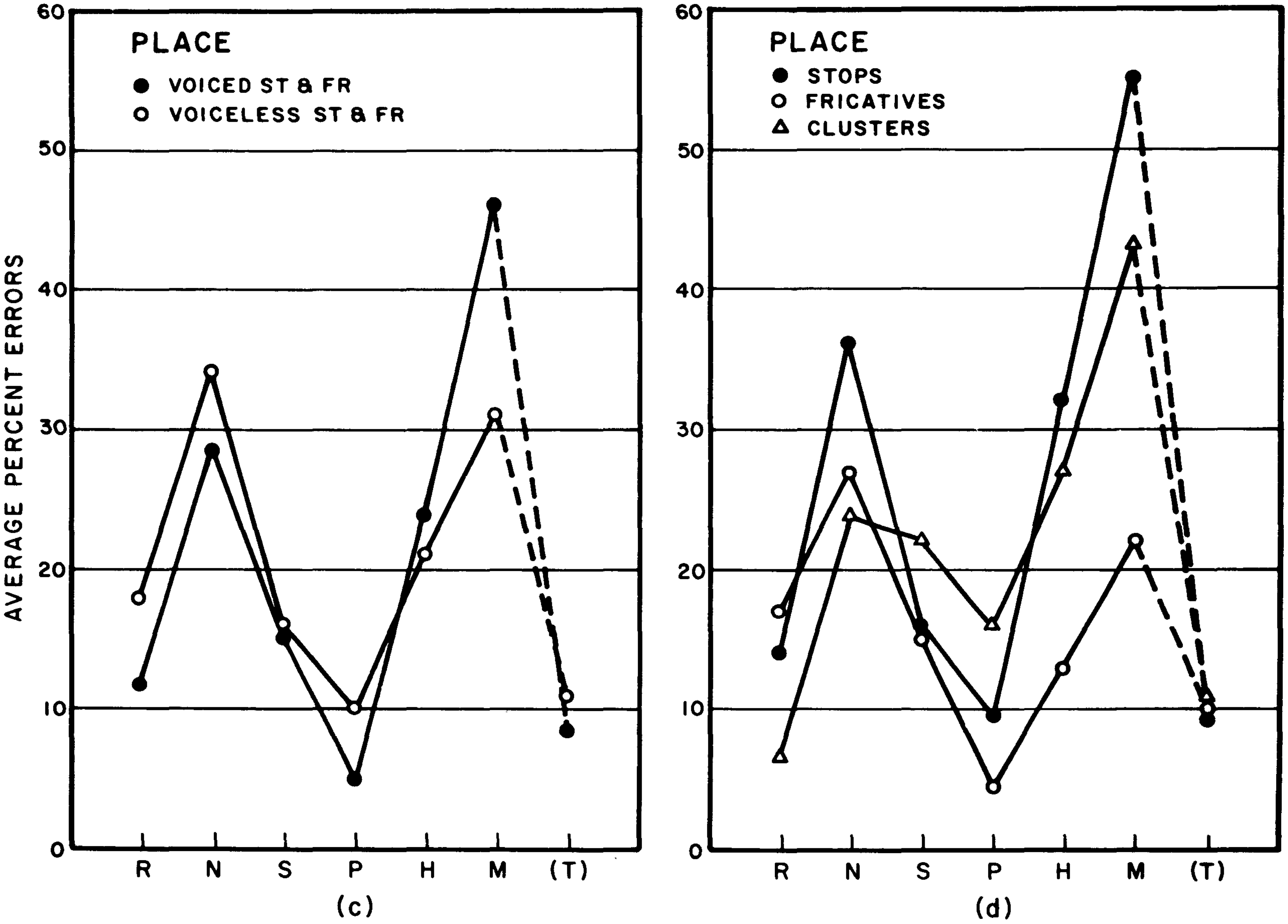

PLACE
● VOICED ST & FR
○ VOICELESS ST & FR
AVERAGE PERCENT ERRORS
60
50
40
30
20
10
0
R
N
S
P
H
M
(T)
(c)
PLACE
● STOPS
○ FRICATIVES
△ CLUSTERS
60
50
40
30
20
10
0
R
N
S
P
H
M
(T)
(d)

FIG. 2.4-3 (CONT.)

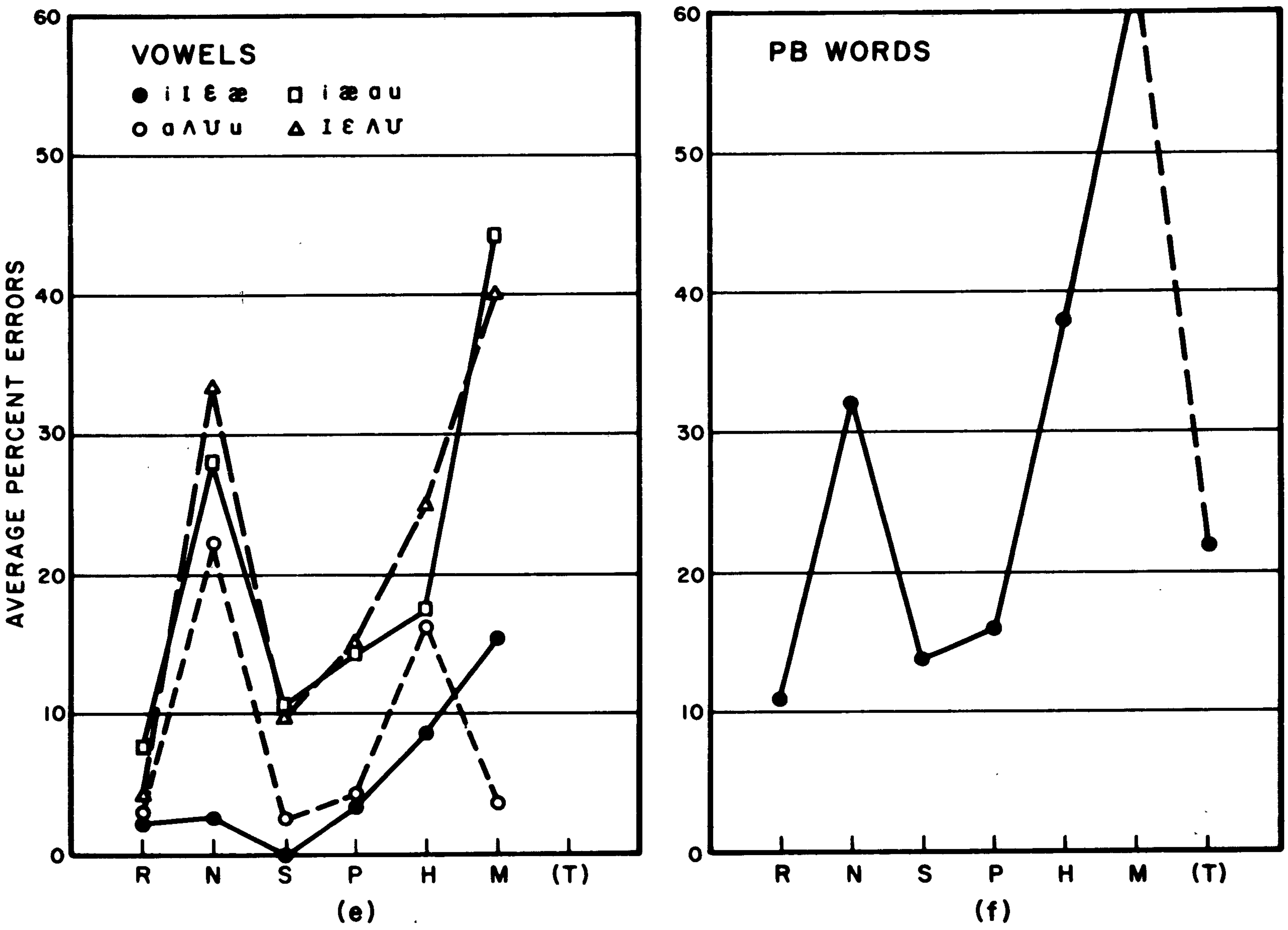

VOWELS
● i ɪ ɛ æ
□ i æ ɑ u
○ ɑ ʌ ʊ u
△ ɪ ɛ ʌ ʊ
AVERAGE PERCENT ERRORS
60
50
40
30
20
10
0
R
N
S
P
H
M
(T)
(e)
PB WORDS
60
50
40
30
20
10
0
R
N
S
P
H
M
(T)
(f)

FIG. 2.4-3 (CONT.)

Rationale for interpretation of nonsense syllable tests. The interpretation of the data from the nonsense syllable tests may be facilitated by a brief review of the primary acoustic features that seem to signal the various distinctions among the vowels and consonants of natural speech. A given speech compression technique may introduce distortion of some of these features but may leave others relatively unchanged. Thus, in order to interpret the confusion matrices that are obtained for a given system, it is necessary to know the temporal and spectral properties of the speech sounds that form the major cues for identification of the sounds.

The types of acoustic features that signal place of production for a consonant depend greatly on the manner of production, i.e., on the row in which the consonant appears in Table 2.4-1. The cues for place of production of some of the consonants are carried by the properties of the transitions to and from the adjacent vowels. It is known, for example, that the voiced stops /b d g/ are distinguished from each other largely on the basis of the transitions of the formants in the adjacent vowel, particularly the second formant.[5,7,29,30] The same is true of the nasal consonants,[5,29,30] the two voiceless fricatives /f θ/[20,23] and, to some extent, the voiceless stops.[15] Some of the consonants, however, such as the voiceless fricatives /f s ʃ/, the voiced fricatives /v z ʒ /, and to some extent the voiceless stops, are undoubtedly identified on the basis of their spectral characteristics during intervals when the short-time spectrum is relatively stationary.[20,23] The liquids and glides are characterized by changing formants, although there are apparently certain approximate target positions for the first two or three formants of these sounds.[42] The glide /w/, for example, is characterized by two low-frequency resonances (roughly 250 and 700 cps for male voices), while /j/ is characterized by a low-frequency resonance (about 250 cps) and two closely-spaced resonances at high frequencies (around 2500-3000 cps).

The identification of a vowel is known to depend primarily on the frequency locations F_1 and F_2 of the lowest two vocal-tract resonances or formants.[43] These formants are manifested in the acoustic spectrum by spectral peaks whose frequency locations and relative amplitudes depend upon the frequencies of the formants. Approximate average values of the first two formant frequencies at centrally located points in the vowels for the talkers who recorded the nonsense syllable tests are given in Table 2.4-14.[24] From this table it is observed that vowels in the front vowel series /i ɪ ɛ æ/ are characterized by successively increasing values of F_1 and successively decreasing values of F_2. For the back vowel series /ɑ ʌ ʊ u/ the first two formants are relatively close together and the frequency of the first formant progressively decreases. In both back and front vowel series, vowel durations as well as formant frequencies can provide cues to identification of the vowel.

The long vowels /i æ ɑ u/ and the short vowels /ɪ ɛ ʌ ʊ/ can be grouped into the pairs /i u/, /æ ɑ/, ,ɪ ʊ/, and /ɛ ʌ/. Within each pair the frequency of the first formant is about the same, and there is a difference only in the frequency of the second formant. Thus, any processing that results in distortion or attenuation of the spectrum in the range of the second formant frequency may lead to confusions within these pairs of vowels.

Table 2.4-4. **RESULTS OF NONSENSE SYLLABLE TESTS**

Reference System

1

	VLS	VD
VLS	96.8	3.2
VD	2.8	97.2

2

	ST	FR
ST	93.5	6.5
FR	8.9	91.1

VOICELESS

3

	ST	FR
ST	94.5	5.5
FR	7.2	92.8

VOICED

4

	ST	FR	NA
ST	92.2	6.1	1.7
FR	8.2	87.4	4.4
NA	0.6	1.3	98.1

5

	GL	NA	LI
GL	93.9	0.9	5.2
NA	0.5	96.5	3.0
LI	2.6	0.5	96.9

INITIAL

6

	NA	LI
NA	100	--
LI	--	100

FINAL

Table 2.4-4. (cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Reference System)

7

	f	s	ʃ
f	88.5	11.1	0.4
s	35.6	61.9	2.5
ʃ	1.8	14.4	83.8

8

	f	θ	s	ʃ
f	81.8	13.0	4.8	0.4
θ	31.0	39.6	28.9	0.5
s	13.6	12.1	72.6	1.7
ʃ	1.4	2.1	14.4	82.1

9

	p	t	k
p	89.1	6.2	4.7
t	7.4	86.8	5.8
k	3.8	16.5	79.7

10

	v	z	ʒ
v	93.2	6.4	0.4
z	14.7	82.1	3.2
ʒ	0.7	9.4	89.9

11

	v	z
v	90.4	9.6
z	7.5	92.5

12

	b	d	g
b	87.3	11.3	1.4
d	4.1	92.0	3.9
g	3.9	14.2	81.9

13

	b	d
b	95.2	4.8
d	2.4	97.6

14

	m	n
m	94.3	5.7
n	6.3	93.7

INITIAL & FINAL

15

	m	n	ŋ
m	79.9	20.1	--
n	3.5	94.8	1.7
ŋ	5.5	29.2	65.3

FINAL

16

	r	ℓ
r	97.2	2.8
ℓ	0.9	99.1

INITIAL

17

	r	ℓ
r	99.4	0.6
ℓ	0.6	99.4

FINAL

18

	w	j
w	100	--
j	12.7	87.3

Table 2.4-4. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Reference System)

19

	s	sp	st	sw	sℓ	sm	sn	str
s	98.3	--	0.9	--	--	0.8	--	--
sp	--	98.3	1.7	--	--	--	--	--
st	0.9	4.2	92.6	--	1.5	--	--	0.8
sw	--	--	--	100	--	--	--	--
sℓ	6.2	--	0.8	1.8	87.8	3.4	--	--
sm	--	--	--	--	2.7	80.5	16.8	--
sn	--	--	--	0.9	5.4	13.7	80.0	--
str	--	--	--	--	--	--	--	100

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	83.1	4.1	--	--	--	0.8	11.2	0.8	--
t	--	96.9	--	--	--	2.3	0.8	--	--
n	--	--	99.2	--	--	--	--	--	0.8
ℓ	--	--	--	99.1	0.9	--	--	--	--
r	--	--	0.9	--	99.1	--	--	--	--
rt	--	1.6	--	--	--	98.4	--	---	--
st	8.6	0.8	--	--	--	--	90.6	--	--
ℓt	--	4.4	--	--	--	0.9	1.8	92.9	--
nt	--	1.8	--	--	--	--	--	--	98.2

Table 2.4-4. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Reference System)

21

	i	ɪ	ɛ	æ
i	96.9	2.3	0.8	--
ɪ	--	98.4	--	1.6
ɛ	--	--	99.2	0.8
æ	0.8	1.6	0.8	96.8

22

	ɑ	ʌ	ʊ	u
ɑ	99.2	0.8	--	--
ʌ	--	98.4	0.8	0.8
ʊ	0.8	2.3	94.6	2.3
u	--	0.8	3.1	96.1

23

	i	æ	ɑ	u
i	96.1	--	--	3.9
æ	--	88.3	11.7	--
ɑ	--	14.8	85.2	--
u	--	--	--	100

24

	ɪ	ɛ	ʌ	ʊ
ɪ	100	--	--	--
ɛ	--	98.5	1.5	--
ʌ	--	6.6	92.7	0.7
ʊ	4.4	--	3.7	91.9

Table 2.4-5. **RESULTS OF NONSENSE SYLLABLE TESTS**

Narrow-Band System

1

	VLS	VD
VLS	98.0	2.0
VD	2.8	97.2

2

	ST	FR
ST	95.6	4.4
FR	1.2	98.8

VOICELESS

3

	ST	FR
ST	96.0	4.0
FR	7.4	92.6

VOICED

4

	ST	FR	NA
ST	95.6	1.9	2.5
FR	15.5	78.3	6.2
NA	0.7	6.5	92.8

5

	GL	NA	LI
GL	61.5	18.4	20.1
NA	2.1	93.1	4.8
LI	24.6	30.5	44.5

INITIAL

6

	NA	LI
NA	95.7	4.5
LI	14.1	85.9

FINAL

Table 2.4-5. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Narrow-Band System)

7

	f	s	ʃ
f	35.1	64.0	0.9
s	13.8	84.6	1.6
ʃ	0.4	4.1	95.5

8

	f	θ	s	ʃ
f	23.9	16.0	52.8	7.3
θ	25.7	27.5	45.8	1.0
s	8.0	4.5	79.5	8.0
ʃ	1.1	--	4.8	94.1

9

	p	t	k
p	40.7	56.0	3.3
t	1.1	93.2	5.7
k	5.0	48.4	46.6

10

	v	z	ʒ
v	40.4	58.9	0.7
z	11.5	85.8	2.7
ʒ	0.4	3.8	95.8

11

	v	z
v	35.0	65.0
z	5.0	95.0

12

	b	d	g
b	65.4	30.8	3.8
d	1.5	94.0	4.5
g	2.9	51.1	46.0

13

	b	d
b	69.2	30.8
d	2.8	97.2

14

	m	n
m	45.0	55.0
n	14.0	86.0

INITIAL & FINAL

15

	m	n	ŋ
m	14.1	83.0	2.9
n	5.3	82.8	11.9
ŋ	9.2	40.8	50.0

FINAL

16

	r	ℓ
r	65.5	34.5
ℓ	6.9	93.1

INITIAL

17

	r	ℓ
r	60.2	39.8
ℓ	10.3	89.7

FINAL

18

	w	j
w	97.5	2.5
j	8.1	91.9

Table 2.4-5. (Cont.) **RESULTS OF NONSENSE SYLLABLE TESTS** (Narrow-Band System)

19

	s	sp	st	sw	sℓ	sm	sn	str
s	95.2	0.7	2.0	0.7	1.4	--	--	--
sp	--	84.7	13.9	0.7	--	0.7	--	--
st	1.4	9.0	85.4	--	--	0.7	--	3.5
sw	4.9	0.7	0.7	78.8	11.1	0.7	2.1	--
sℓ	9.0	0.7	--	4.2	69.5	5.5	10.4	0.7
sm	--	2.8	--	9.7	4.1	63.9	18.8	0.7
sn	--	0.7	--	5.5	15.3	33.4	45.1	---
str	--	0.7	9.0	--	0.7	--	--	89.6

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	78.4	2.8	--	0.7	--	--	13.9	2.8	1.4
t	--	84.0	--	--	--	3.5	1.4	4.9	6.3
n	--	--	96.5	0.7	--	0.7	0.7	--	1.4
ℓ	1.4	0.7	2.1	86.6	7.6	--	--	0.7	0.7
r	1.4	--	6.2	47.3	44.4	0.7	--	--	--
rt	--	12.5	--	--	--	64.6	0.7	20.1	2.1
st	11.1	2.8	--	--	--	--	85.4	--	0.7
ℓt	--	4.8	--	--	--	10.4	2.1	76.5	6.2
nt	--	18.1	0.7	--	--	4.8	1.4	3.5	71.5

Table 2.4-5. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Narrow-Band System)

21

	i	I	ɛ	æ
i	95.8	3.5	0.7	--
I	0.7	99.3	--	--
ɛ	0.7	1.4	97.9	--
æ	--	0.7	2.8	96.5

22

	ɑ	ʌ	ʊ	u
ɑ	93.8	4.8	--	1.4
ʌ	1.4	93.1	4.8	0.7
ʊ	2.1	52.1	44.4	1.4
u	0.7	4.8	13.9	80.6

23

	i	æ	ɑ	u
i	96.5	--	1.4	2.1
æ	0.7	72.9	26.4	--
ɑ	--	62.5	36.8	0.7
u	15.3	1.4	1.4	81.9

24

	I	ɛ	ʌ	ʊ
I	92.4	--	--	7.6
ɛ	0.7	82.6	15.3	1.4
ʌ	--	50.7	49.3	--
ʊ	56.3	0.7	0.7	42.3

Table 2.4-6. RESULTS OF NONSENSE SYLLABLE TESTS

Semi-vocoder

1

	VLS	VD
VLS	98.4	1.6
VD	1.1	98.9

2

	ST	FR
ST	94.5	5.5
FR	7.8	92.2

VOICELESS

3

	ST	FR
ST	98.2	1.8
FR	8.2	91.8

VOICED

4

	ST	FR	NA
ST	92.9	5.0	2.1
FR	11.9	80.3	7.8
NA	1.4	1.7	96.9

5

	GL	NA	LI
GL	73.1	10.1	16.8
NA	3.4	93.8	2.8
LI	7.7	1.2	91.1

INITIAL

6

	NA	LI
NA	99.2	0.8
LI	1.0	99.0

FINAL

Table 2.4-6. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Semi-vocoder)

7

	f	s	ʃ
f	92.0	7.4	0.6
s	39.8	57.4	2.8
ʃ	0.4	1.2	98.4

8

	f	θ	s	ʃ
f	73.5	21.4	4.7	0.4
θ	53.0	36.4	10.2	0.4
s	24.5	25.8	42.8	6.9
ʃ	--	0.4	1.1	98.5

9

	p	t	k
p	83.8	7.2	9.0
t	5.0	84.0	11.0
k	3.7	9.0	87.3

10

	v	z	ʒ
v	89.6	9.6	0.8
z	21.5	77.0	1.5
ʒ	--	1.8	98.2

11

	v	z
v	81.2	18.8
z	10.1	89.9

12

	b	d	g
b	79.3	14.5	6.2
d	4.3	83.4	12.3
g	3.1	11.5	85.4

13

	b	d
b	84.2	15.8
d	4.0	96.0

14

	m	n
m	85.4	14.6
n	11.0	89.0

INITIAL & FINAL

15

	m	n	ŋ
m	57.9	40.5	1.6
n	3.4	90.4	6.2
ŋ	3.8	21.3	74.9

FINAL

16

	r	ℓ
r	87.8	12.2
ℓ	7.2	92.8

INITIAL

17

	r	ℓ
r	100	--
ℓ	1.5	98.5

FINAL

18

	w	j
w	98.7	1.3
j	--	100

Table 2.4-6. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Semi-vocoder)

19

	s	sp	st	sw	sℓ	sm	sn	str
s	78.7	2.9	5.8	1.5	2.1	1.4	3.1	4.5
sp	5.3	81.9	10.7	1.4	--	0.8	--	--
st	0.8	13.6	78.2	--	0.8	--	--	6.6
sw	1.5	0.8	--	82.0	12.0	0.8	0.7	2.2
sℓ	8.5	1.5	1.4	5.8	68.0	5.1	5.1	4.6
sm	--	--	--	0.7	12.5	52.1	34.7	--
sn	--	--	--	0.7	15.1	24.1	60.1	--
str	--	--	2.2	0.7	0.7	--	--	96.4

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	51.7	23.4	0.8	--	0.8	0.8	21.0	0.8	0.7
t	0.8	83.8	--	--	--	5.2	3.8	0.8	5.4
n	0.7	--	97.1	1.4	0.8	--	--	--	--
ℓ	--	--	0.7	97.1	2.2	--	--	--	--
r	--	--	--	8.6	91.4	--	--	--	--
rt	--	2.3	--	--	1.5	96.2	--	--	--
st	15.0	6.6	--	--	0.8	5.9	68.7	0.7	2.3
ℓt	--	5.9	5.4	--	--	12.5	0.7	73.4	2.1
nt	--	12.1	--	--	--	0.8	--	--	87.1

Table 2.4-6. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Semi-vocoder)

21

	i	I	ɛ	æ
i	100	--	--	--
I	--	100	--	--
ɛ	--	--	100	--
æ	--	--	--	100

22

	ɑ	ʌ	ʊ	u
ɑ	99.2	0.8	--	--
ʌ	--	97.6	0.8	1.6
ʊ	--	3.9	96.1	--
u	--	--	1.6	98.4

23

	i	æ	ɑ	u
i	96.9	--	0.8	2.3
æ	0.8	74.2	25.0	--
ɑ	--	9.4	90.6	--
u	2.3	--	—	97.7

24

	I	ɛ	ʌ	ʊ
I	93.7	--	--	6.3
ɛ	--	98.4	0.8	0.8
ʌ	--	9.4	90.6	--
ʊ	18.7	0.8	1.6	78.9

Table 2.4-7. RESULTS OF NONSENSE SYLLABLE TESTS

Philco Vocoder

1

	VLS	VD
VLS	95.3	4.7
VD	2.4	97.6

2

	ST	FR
ST	96.8	3.2
FR	12.8	87.2

VOICELESS

3

	ST	FR
ST	85.6	14.4
FR	5.3	94.7

VOICED

4

	ST	FR	NA
ST	80.4	17.5	2.1
FR	9.7	88.3	2.0
NA	1.3	3.9	94.8

5

	GL	NA	LI
GL	79.5	2.8	17.7
NA	4.5	90.6	4.9
LI	7.6	7.3	85.1

INITIAL

6

	NA	LI
NA	98.9	1.1
LI	1.2	98.8

FINAL

Table 2.4-7. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Philco Vocoder)

7

	f	s	ʃ
f	94.6	4.4	1.0
s	4.0	93.3	2.7
ʃ	0.4	5.1	94.5

8

	f	θ	s	ʃ
f	78.1	17.7	4.2	--
θ	56.6	41.7	1.0	0.7
s	8.4	10.4	81.2	--
ʃ	--	--	1.4	98.6

9

	p	t	k
p	77.9	9.2	12.9
t	2.1	90.6	7.3
k	7.1	2.8	90.1

10

	v	z	ʒ
v	96.5	3.3	0.2
z	3.0	95.2	1.8
ʒ	0.2	1.9	97.9

11

	v	z
v	97.3	2.7
z	3.3	96.7

12

	b	d	g
b	75.3	7.7	17.0
d	3.3	84.8	11.9
g	4.5	7.0	88.5

13

	b	d
b	88.9	11.1
d	0.5	99.5

14

	m	n
m	70.2	29.8
n	10.0	90.0

INITIAL & FINAL

15

	m	n	ŋ
m	30.3	52.8	16.9
n	4.6	84.1	11.3
ŋ	3.8	30.9	75.3

FINAL

16

	r	ℓ
r	95.1	4.9
ℓ	12.4	87.6

INITIAL

17

	r	ℓ
r	99.5	0.5
ℓ	2.9	97.1

FINAL

18

	w	j
w	100	--
j	0.8	99.2

Table 2.4-7. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Philco Vocoder)

19

	s	sp	st	sw	sℓ	sm	sn	str
s	95.7	1.4	--	0.7	2.1	--	--	--
sp	1.4	77.1	19.4	--	0.7	--	1.4	--
st	--	4.2	88.9	--	--	--	--	6.9
sw	--	--	--	91.6	4.9	2.1	1.4	--
sℓ	21.5	--	--	2.8	61.1	2.1	8.3	4.2
sm	--	--	0.7	4.8	1.4	50.1	43.0	--
sn	0.7	--	--	0.7	2.1	6.3	89.5	0.7
str	--	--	--	0.7	--	--	--	99.3

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	63.9	14.6	--	--	--	3.4	17.4	--	0.7
t	0.7	77.8	--	--	--	1.4	2.1	11.1	6.9
n	--	--	97.2	--	2.1	--	--	--	0.7
ℓ	--	--	--	98.6	0.7	--	--	0.7	--
r	--	--	--	0.7	98.6	0.7	--	--	--
rt	0.7	2.8	--	--	--	92.3	--	1.4	2.8
st	4.9	7.6	--	--	--	--	86.8	--	0.7
ℓt	--	1.4	6.2	7.7	--	4.9	--	76.4	3.4
nt	--	0.7	2.7	--	--	1.4	--	9.1	86.1

Table 2.4-7. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Philco Vocoder)

21

	i	I	ɛ	æ
i	97.6	1.6	--	0.8
I	0.8	96.7	2.5	--
ɛ	--	--	95.8	4.2
æ	--	0.8	1.6	97.6

22

	ɑ	ʌ	ʊ	u
ɑ	100	--	--	--
ʌ	4.2	93.3	2.5	--
ʊ	--	3.8	93.7	2.5
u	--	0.8	3.8	95.8

23

	i	æ	ɑ	u
i	98.4	--	--	1.6
æ	--	68.3	31.7	--
ɑ	--	23.3	76.7	--
u	--	--	--	100

24

	I	ɛ	ʌ	ʊ
I	85.9	2.5	0.8	10.8
ɛ	2.5	85.0	10.7	0.8
ʌ	--	14.2	84.2	1.6
ʊ	5.9	1.6	2.5	90.0

Table 2.4-8. **RESULTS OF NONSENSE SYLLABLE TESTS**

Hughes Vocoder

1

	VLS	VD
VLS	90.9	9.1
VD	4.7	95.3

2

	ST	FR
ST	80.5	19.5
FR	19.3	80.7

VOICELESS

3

	ST	FR
ST	75.7	24.3
FR	27.7	72.3

VOICED

4

	ST	FR	NA
ST	64.6	30.1	5.3
FR	22.8	72.1	5.1
NA	5.0	7.9	87.1

5

	GL	NA	LI
GL	64.3	30.2	5.5
NA	0.4	92.4	7.2
LI	18.1	20.5	61.4

INITIAL

6

	NA	LI
NA	99.7	0.3
LI	2.2	97.8

FINAL

Table 2.4-8. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Hughes Vocoder)

7

	f	s	ʃ
f	89.1	9.8	1.1
s	13.8	79.9	6.3
ʃ	0.4	3.9	95.7

8

	f	θ	s	ʃ
f	65.3	34.0	0.7	--
θ	65.3	32.9	1.8	--
s	14.6	10.1	72.7	2.6
ʃ	--	0.7	2.7	96.6

9

	p	t	k
p	62.7	21.4	15.9
t	8.2	9.8	82.0
k	16.5	18.4	65.1

10

	v	z	ʒ
v	97.1	2.9	--
z	28.0	68.8	3.2
ʒ	6.6	3.2	90.2

11

	v	z
v	96.9	3.1
z	18.2	81.8

12

	b	d	g
b	60.0	25.8	14.2
d	9.8	75.7	14.5
g	8.2	26.5	65.3

13

	b	d
b	62.3	37.7
d	6.4	93.6

14

	m	n
m	85.7	14.3
n	32.7	67.3

INITIAL & FINAL

15

	m	n	ŋ
m	57.9	34.2	7.9
n	21.0	70.4	8.6
ŋ	16.6	29.2	54.2

FINAL

16

	r	ℓ
r	94.6	5.4
ℓ	8.0	92.0

INITIAL

17

	r	ℓ
r	78.9	21.1
ℓ	1.1	98.9

FINAL

18

	w	j
w	97.1	2.9
j	--	100

Table 2.4-8. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Hughes Vocoder)

19

	s	sp	st	sw	sℓ	sm	sn	str
s	94.0	--	--	1.5	3.8	0.7	--	--
sp	--	82.6	17.4	--	--	--	--	--
st	0.7	90.6	--	--	--	--	0.8	3.9
sw	--	1.6	2.3	78.8	12.6	0.8	0.7	3.2
sℓ	16.2	0.8	3.9	5.0	64.6	1.5	7.9	--
sm	1.4	--	0.8	14.8	9.9	52.7	19.6	0.8
sn	--	--	0.7	3.9	12.5	21.3	61.6	--
str	0.8	6.2	7.0	--	--	0.8	--	85.2

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	48.6	11.1	--	7.7	3.5	4.9	15.2	9.0	--
t	3.6	37.4	3.5	2.1	2.8	13.5	0.7	24.0	12.4
n	--	--	98.5	--	0.8	--	--	--	0.7
ℓ	--	--	--	97.2	1.4	--	--	0.7	0.7
r	--	0.7	5.3	30.2	63.1	0.7	--	--	--
rt	1.4	--	0.7	8.3	10.4	63.9	0.7	10.8	3.8
st	17.3	15.4	1.4	2.8	--	6.4	42.7	11.7	1.4
ℓt	--	--	7.0	29.1	0.7	--	--	62.5	0.7
nt	--	3.0	31.6	1.5	--	0.8	--	10.6	52.5

Table 2.4-8. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Hughes Vocoder)

21

	i	I	ɛ	æ
i	97.7	1.5	0.8	--
I	3.9	84.4	11.7	--
ɛ	--	7.0	85.2	7.8
æ	--	--	1.5	98.5

22

	ɑ	ʌ	ʊ	u
ɑ	93.0	5.5	1.5	--
ʌ	6.3	82.8	9.4	1.5
ʊ	0.8	9.4	85.9	3.9
u	--	--	6.3	93.7

23

	i	æ	ɑ	u
i	90.6	--	--	9.4
æ	1.5	62.6	35.9	--
ɑ	--	11.7	87.5	0.8
u	8.6	0.8	0.8	89.8

24

	I	ɛ	ʌ	ʊ
I	71.8	14.1	0.8	13.3
ɛ	3.1	66.4	25.8	4.7
ʌ	--	11.7	79.7	8.6
ʊ	7.8	3.9	5.5	82.8

Table 2.4-9. RESULTS OF NONSENSE SYLLABLE TESTS

Melpar System

1

	VLS	VD
VLS	89.8	10.2
VD	12.0	88.0

2

	ST	FR
ST	73.5	26.5
FR	15.6	84.4

VOICELESS

3

	ST	FR
ST	66.4	33.6
FR	29.3	70.7

VOICED

4

	ST	FR	NA
ST	47.6	37.1	15.3
FR	22.5	65.2	12.3
NA	22.3	28.7	49.0

5

	GL	NA	LI
GL	24.3	18.0	57.8
NA	22.2	41.8	36.0
LI	18.8	7.4	73.8

INITIAL

6

	NA	LI
NA	81.9	18.1
LI	15.4	84.6

FINAL

Table 2.4-9. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Melpar System)

7

	f	s	ʃ
f	87.0	8.0	5.0
s	13.5	83.0	3.5
ʃ	3.7	5.1	91.2

8

	f	θ	s	ʃ
f	69.1	16.5	12.5	1.9
θ	36.4	24.6	35.1	3.9
s	0.8	1.2	96.8	1.2
ʃ	6.3	0.8	7.0	85.9

9

	p	t	k
p	65.9	23.7	10.4
t	28.0	61.9	10.1
k	49.5	25.4	25.3

10

	v	z	ʒ
v	77.6	13.6	8.8
z	26.3	63.1	10.1
ʒ	17.1	18.5	64.4

11

	v	z
v	82.4	17.6
z	17.4	82.6

12

	b	d	g
b	38.2	11.9	49.9
d	27.9	28.2	43.9
g	25.9	22.2	51.9

13

	b	d
b	53.3	46.7
d	42.6	57.4

14

	m	n
m	59.6	40.4
n	50.6	49.4

INITIAL & FINAL

15

	m	n	ŋ
m	23.7	54.6	21.7
n	25.5	51.8	22.7
ŋ	19.2	45.9	34.9

FINAL

16

	r	ℓ
r	26.2	73.8
ℓ	26.1	73.9

INITIAL

17

	r	ℓ
r	52.7	47.3
ℓ	23.1	76.9

FINAL

18

	w	j
w	82.3	27.7
j	54.5	45.5

Table 2.4-9. (Cont.) RESULTS OF NONSENSE SYLLABLE TESTS (Melpar System)

19

	s	sp	st	sw	sℓ	sm	sn	str
s	90.7	0.8	--	3.1	3.8	--	1.6	--
sp	3.1	57.0	19.5	10.2	4.7	0.8	3.9	0.8
st	5.5	6.2	79.8	1.5	1.5	0.8	--	4.7
sw	13.3	--	1.5	62.5	18.3	0.8	3.1	--
sℓ	27.3	0.8	--	21.1	40.7	4.6	3.9	1.6
sm	0.8	0.8	0.8	24.2	47.7	10.9	14.8	--
sn	7.8	0.8	2.3	14.8	42.2	18.8	12.5	0.8
str	1.6	6.3	37.4	0.8	1.5	0.8	--	51.6

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	93.0	0.8	--	--	0.8	1.6	3.8	--	--
t	3.8	74.3	0.8	1.5	--	0.8	2.3	12.6	3.9
n	1.6	0.8	86.6	7.0	1.6	--	0.8	0.8	0.8
ℓ	--	0.8	--	80.6	15.2	1.6	--	0.8	--
r	1.6	--	3.9	51.5	41.4	--	1.6	--	--
rt	4.8	10.1	--	1.6	2.3	23.4	0.8	51.6	5.4
st	43.7	1.6	--	--	--	--	53.9	--	0.8
ℓt	--	7.0	2.3	10.9	--	10.9	5.5	60.2	3.2
nt	5.4	8.6	16.5	0.7	1.6	7.0	--	14.8	45.4

Table 2.4-9. (Cont.)RESULTS OF NONSENSE SYLLABLE TESTS (Melpar System)

21

	i	ɪ	ɛ	æ
i	83.0	9.6	5.9	1.5
ɪ	5.1	80.9	8.1	5.9
ɛ	0.7	2.2	82.4	14.7
æ	--	2.2	5.1	92.7

22

	ɑ	ʌ	ʊ	u
ɑ	98.5	1.5	--	--
ʌ	2.2	97.8	--	--
ʊ	0.7	4.4	94.9	--
u	1.5	0.7	3.7	94.1

23

	i	æ	ɑ	u
i	29.4	--	--	70.6
æ	--	25.0	75.0	--
ɑ	--	25.0	75.0	--
u	5.2	--	0.7	94.1

24

	ɪ	ɛ	ʌ	ʊ
ɪ	32.4	0.7	1.5	65.4
ɛ	3.7	38.2	55.1	3.0
ʌ	--	14.7	85.3	--
ʊ	15.4	1.5	--	83.1

Table 2.4-10. **RESULTS OF NONSENSE SYLLABLE TESTS**

Results of restricted set of nonsense syllable tests for Tasaroff-Daguet system

1

	VLS	VD
VLS	96.7	3.3
VD	2.7	97.3

2

	ST	FR
ST	96.5	3.5
FR	2.4	97.6

VOICELESS

3

	ST	FR
ST	100	--
FR	13.2	86.8

VOICED

4 - 5

	NA	GL	LI	FR
NA	98.5	--	--	1.5
GL	6.0	88.5	1.5	4.0
LI	6.0	--	92.5	1.5
FR	--	1.5	4.5	94.0

INITIAL

4 - 6

	NA	FR	LI
NA	98.0	2.0	--
FR	1.0	99.0	--
LI	1.5	--	98.5

FINAL

Table 2.4-10. (Cont.) **RESULTS OF NONSENSE SYLLABLE TESTS** (Tasaroff-Daguet)

7

	f	s	ʃ
f	74.8	22.0	3.2
s	2.5	93.5	4.0
ʃ	--	2.0	98.0

8

	f	θ	s	ʃ
f	62.5	16.5	19.5	1.5
θ	7.0	54.0	36.0	3.0
s	3.0	1.5	76.0	19.5
ʃ	1.5	--	4.0	94.5

9

	p	t	k
p	94.4	2.5	3.1
t	2.4	85.7	11.9
k	2.9	6.9	90.2

10

	v	z	ʒ
v	91.5	6.7	1.8
z	6.2	83.3	10.5
ʒ	0.8	1.5	97.7

11

	v	z
v		
z		

12

	b	d	g
b	91.7	2.0	6.3
d	--	92.3	7.7
g	1.0	7.7	91.3

13

	b	d
b		
d		

14

	m	n
m	100	--
n	--	100

INITIAL

15

	m	n	ŋ
m	78.0	5.5	16.5
n	--	94.0	6.0
ŋ	5.5	5.5	89.0

FINAL

16

	r	ℓ
r	100	--
ℓ	4.0	96.0

INITIAL

17

	r	ℓ
r	100	--
ℓ	--	100

FINAL

18

	w	j
w	100	--
j	--	100

Table 2.4-10. (Cont.) **RESULTS OF NONSENSE SYLLABLE TESTS** (Tasaroff-Daguet)

19

	s	sk	st	sw	sℓ	sm	sn	str
s	100	--	--	--	--	--	--	--
sk	--	84	13	--	--	--	--	3
st	--	--	94	--	--	--	3	3
sw	--	--	--	100	--	--	--	--
sℓ	--	--	--	6	63	3	28	--
sm	--	--	--	--	--	56	44	--
sn	--	--	--	--	--	3	97	--
str	--	--	6	--	--	--	--	94

20

	s	t	n	ℓ	r	rt	st	ℓt	nt
s	75	--	--	--	--	3	16	--	6
t	--	94	--	--	--	--	--	--	6
n	--	--	100	--	--	--	--	--	--
ℓ	--	--	--	100	--	--	--	--	--
r	--	--	--	--	100	--	--	--	--
rt	--	--	--	--	--	100	--	--	--
st	--	--	--	--	--	--	100	--	--
ℓt	--	--	--	--	--	--	--	62	38
nt	--	--	--	--	--	--	--	--	100

Table 2.4-11
List of Confusion Matrices
For Each of Six Speech-Processing Systems

1. Voiced-voiceless: Tests 1 and 4
2. Interrupted-continuant (voiceless): Test 2
3. Interrupted-continuant (voiced): Test 3
4. Voiced stop-fricative-nasal manner: Test 5
5. Nasal-liquid-glide manner (initial): Test 7
6. Nasal-liquid manner (final): Test 7
7. Voiceless fricatives /f s ʃ/: Tests 2 and 4
8. Voiceless fricatives /f θ s ʃ/: Test 6
9. Voiceless stops /p t k/: Tests 1 and 2
10. Voiced fricatives /v z ʒ/: Tests 3 and 4
11. Voiced fricatives /v z/: Test 5
12. Voiced stops /b d g/: Tests 1 and 3
13. Voiced stops /b d/: Test 5
14. Nasals /m n/ (initial and final): Tests 5 and 7
15. Nasals /m n ŋ/ (final): Test 7
16. Liquids /r ℓ/ (initial): Test 7
17. Liquids /r ℓ/ (final): Test 7
18. Glides /w j/: Test 7
19. Consonant clusters (initial): Test 9
20. Consonant clusters (final): Test 8
21. Vowels /i ɪ ɛ æ/: Test 10
22. Vowels /ɑ ʌ ʊ u/: Test 11
23. Vowels /i æ ɑ u/: Test 12
24. Vowels /ɪ ɛ ʌ ʊ/: Test 13

Table 2.4-12

Probability (times 100) That a Stimulus Is Categorized Incorrectly for Each of the 24 Types of Confusion Listed in Table 2.4-11. The Letters at the Top of Each Column Identify the Various Speech-Processing Systems (See Key in Section 2.1)

	R	N	S	P	H	M	T*
1.	3.0	2.4	1.4	3.6	6.9	11.1	3.0
2.	7.7	2.8	6.7	8.0	19.4	21.0	2.9
3.	6.4	5.7	5.0	9.8	24.6	31.4	6.6
4.	7.4	11.1	10.0	12.2	25.4	46.2	
5.	4.2	33.8	12.2	14.9	27.3	42.2	6.8
6.	0	9.3	0.9	1.2	1.2	16.8	0.8
7.	21.9	28.3	17.4	5.9	11.8	12.9	11.2
8.	31.0	43.7	37.4	25.1	33.1	30.9	28.2
9.	14.8	39.8	15.0	13.8	30.1	49.0	9.9
10.	11.6	26.0	11.7	3.4	14.6	31.6	9.2
11.	8.5	35.0	14.4	3.0	10.6	17.5	19.5
12.	12.9	31.5	17.3	5.5	33.0	60.6	8.2
13.	3.6	16.8	9.9	5.8	22.0	44.6	
14.	5.5	34.5	12.8	19.9	23.5	45.5	
15.	20.0	51.0	25.6	40.1	39.2	63.2	13.0
16.	1.9	20.7	9.7	8.7	6.7	50.0	2.0
17.	0.6	25.0	0.8	1.7	11.1	35.2	0
18.	6.3	5.3	0.6	0.4	1.5	41.1	0
19.	8.3	23.6	27.2	18.7	24.0	49.4	14.0
20.	4.9	23.4	17.1	13.6	30.5	35.8	7.6
21.	2.1	2.6	0	3.1	8.5	15.3	
22.	2.9	22.0	2.3	4.2	16.1	3.7	
23.	7.6	28.0	10.2	14.2	17.4	44.1	
24.	4.2	33.3	9.8	15.0	24.8	40.2	

* Data for the Tasaroff-Daguet system are derived from a restricted series of tests.

Table 2.4-13
Probability (times 100) of Incorrect Identification
of Different Consonant Features, as Derived from the
Results of the Nonsense Syllable Tests
Results Represent Averages of Data in Rows of Table 2.4-12,
as Indicated

Feature	Rows of Table 2.4-12	R	N	S	P	H	M	T*
Voiced-voiceless	1	3.0	2.4	1.4	3.6	6.9	11.1	3.0
Interrupted-continuant	2,3	7.0	4.3	5.8	8.9	22.0	26.2	4.8
Other manner	4,5,6	3.9	18.1	7.7	9.4	18.0	35.1	
Voiced stops + fricatives	10,12	12.2	28.8	14.5	4.5	23.8	46.1	8.7
Voiceless stops + fricatives	7,9	18.3	34.0	16.2	9.8	21.0	31.0	10.6
Stops	9,12	13.8	35.6	16.2	9.6	31.6	54.8	9.1
Fricatives	7,10	16.8	27.2	14.5	4.6	13.2	22.3	10.2
Nasals	14,15	12.7	42.8	19.2	30.0	31.4	54.4	
Liquids + glides	16,17,18	2.9	17.0	3.7	3.6	6.4	42.1	0.7
Clusters	19,20	6.6	23.5	22.2	16.1	27.2	42.6	10.8

* Data for the Tasaroff-Daguet system are derived from a restricted series of tests.

Table 2.4-14

Values of the Frequencies F_1 and F_2 of the First and Second Formants for Vowels Occurring in Nonsense Syllables of the Type Used in the Vowel Tests. Averages Are Taken for the Two Talkers Who Recorded the Tests

Vowel	F_1	F_2
	cps	cps
i	300	2150
ɪ	430	1710
ɛ	530	1700
æ	670	1630
ɑ	690	1200
ʌ	600	1300
ʊ	450	1290
u	290	1230

2.5 Voice Quality Tests

High intelligibility scores, obtained for a particular speech compression system under a given set of conditions, do not necessarily imply that the speech output is natural and that the voice quality is high. A particular system may process speech in such a way as to obscure many of the familiar perceptual cues and introduce a new set of consistent cues which, when learned, will help the listener to distinguish between speech sounds that might have been ambiguous before learning. Listeners who are familiar with the system may achieve surprisingly high intelligibility scores and yet rate the same system as inferior to another equally intelligible system on the basis of naturalness or voice quality.

Three voice quality tests of the paired-comparison type, one for each of three speakers, were recorded and administered to two groups of fifteen students each at Tufts University. These tests consisted of 42 randomized sentence* pairs representing all possible forward and reverse combinations of the seven systems under consideration. Identical sentences were repeated as seldom as possible (although some repetition was necessary because each speaker read only five sentences through each system), and the test material and speakers were not used again for other tests.

Two further voice quality tests of the paired-comparison type were made and given to the students. These tests also consisted of 42

* The sentences were the so-called Harvard Test Sentences.[11]

randomized sentence pairs, but the sentence used was always the same phrase: "Number thirty-six, you will write now" (no PB word was included). The test material was obtained by editing appropriate PB word list recordings.

The listeners were given the following sets of instructions:

For the first group of three tests -- "Each test item consists of two sentences. During the pause following each pair you are to write next to the appropriate item on your answer sheet the number _1_ if you think the voice quality of the _first sentence_ of the pair is better, and the number _2_ if you think the voice quality of the _second sentence_ is better. Make your judgments on the intelligibility and naturalness of the speech. There are 42 items in each test."

For the second group of two tests -- "Each test item consists of the sentence: 'Number thirty-six, you will write now,' spoken twice. During the pause following each pair you are to write next to the appropriate item on your answer sheet the number _1_ if you think the voice quality of the _first sentence_ of the pair is better, and the number _2_ if you think the voice quality of the _second sentence_ is better. Make your judgments on the intelligibility and naturalness of the speech. There are 42 items in each test."

The final results obtained for the first group of tests are given in Table 2.5-1. Each entry represents the number of times that a particular system was preferred to all other systems by 30 subjects for three tests (speakers). A distinction is made as to whether the sentence from the preferred system occurred first or second in each sentence pair. From an examination of the corresponding columns it appears that the responses have essentially no time error; i.e., there is no obvious tendency for preference of the second sentence in the pairs regardless of the system used.

Table 2.5-1
Results of Voice Quality Tests Using Different Sentences, for Three Speakers and Thirty Subjects

System	Preferred Sentence First in Pair	Preferred Sentence Second in Pair	Total Relative Preference
Melpar	57	22	79
Stromberg	372	436	808
Philco	291	253	544
Hughes	201	206	407
Tasaroff-Daguet	299	290	589
Narrow-Band	236	217	453
Reference	421	472	893
No. of blanks			7
No. of opportunities to respond			3780

The relative preference for the seven systems is also shown in graphical form in Fig. 2.5-1. This figure was prepared from the right-hand column of Table 2.5-1.

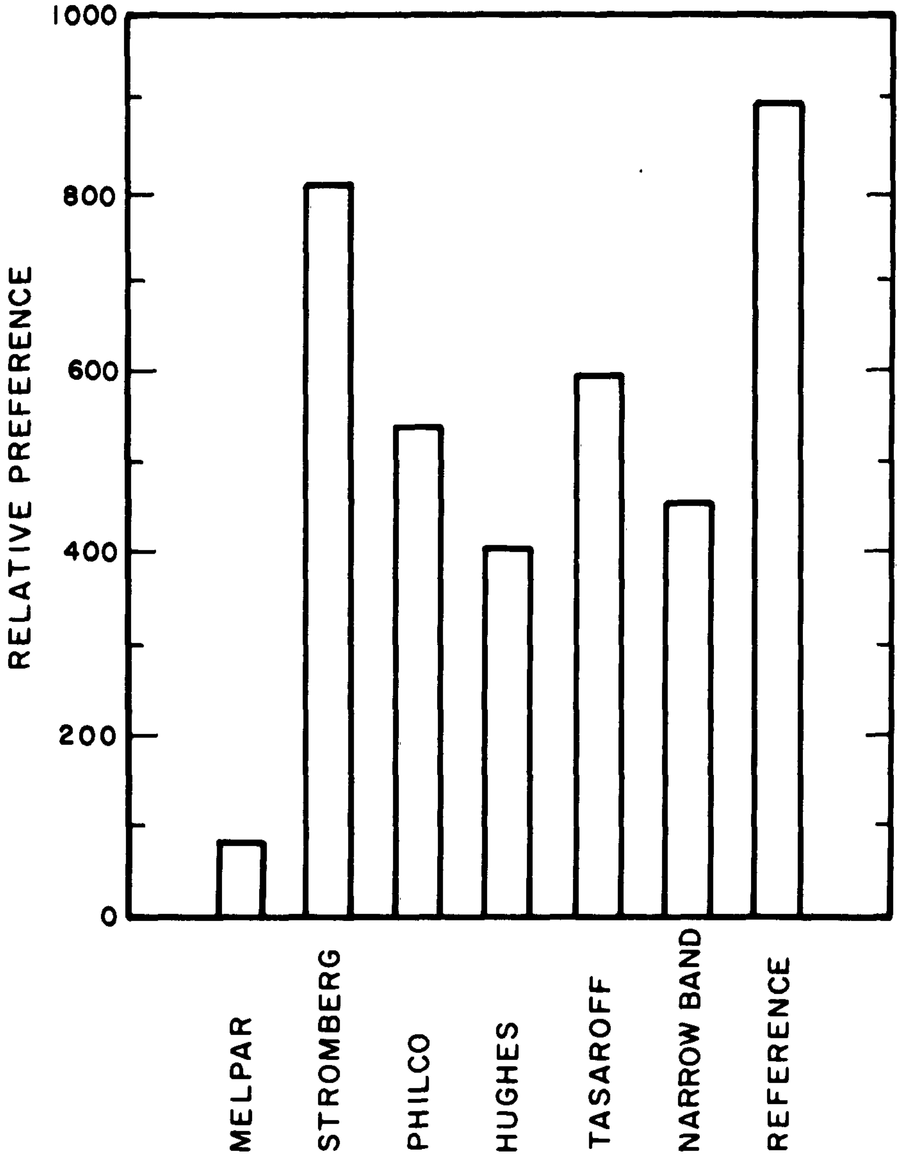


FIG. 2.5-1 RESULTS OF QUALITY TESTS WITH DIFFERENT TEST SENTENCES

The final results obtained for the second group of two tests (using the phrase "Number thirty-six, you will write now") are given in Table 2.5-2. Again each entry represents the number of times that a particular system was preferred to all other systems by 30 subjects for two tests (speakers). In this case there is a slight indication of time error in the responses.

The relative preference for the systems on the basis of voice quality tests using the PB carrier phrase is shown in graphical form in Fig. 2.5-2. This figure was prepared from the right-hand column of Table 2.5-2.

It will be noted that, except for a change between the orders of the scores for the Philco and Tasaroff-Daguet systems, the various systems tested are rank ordered in the same way for both tests of voice quality.

Table 2.5-2

Results of Voice Quality Tests Using the Same Sentence (PB Carrier Phrase), for Two Speakers and Thirty Subjects

System	Preferred Sentence First in Pair	Preferred Sentence Second in Pair	Total Relative Preference
Melpar	19	24	43
Stromberg	203	260	463
Philco	175	225	400
Hughes	100	117	217
Tasaroff-Daguet	173	215	388
Narrow-Band	165	199	364
Reference	303	341	644
No. of blanks			1
No. of opportunities to respond			2520

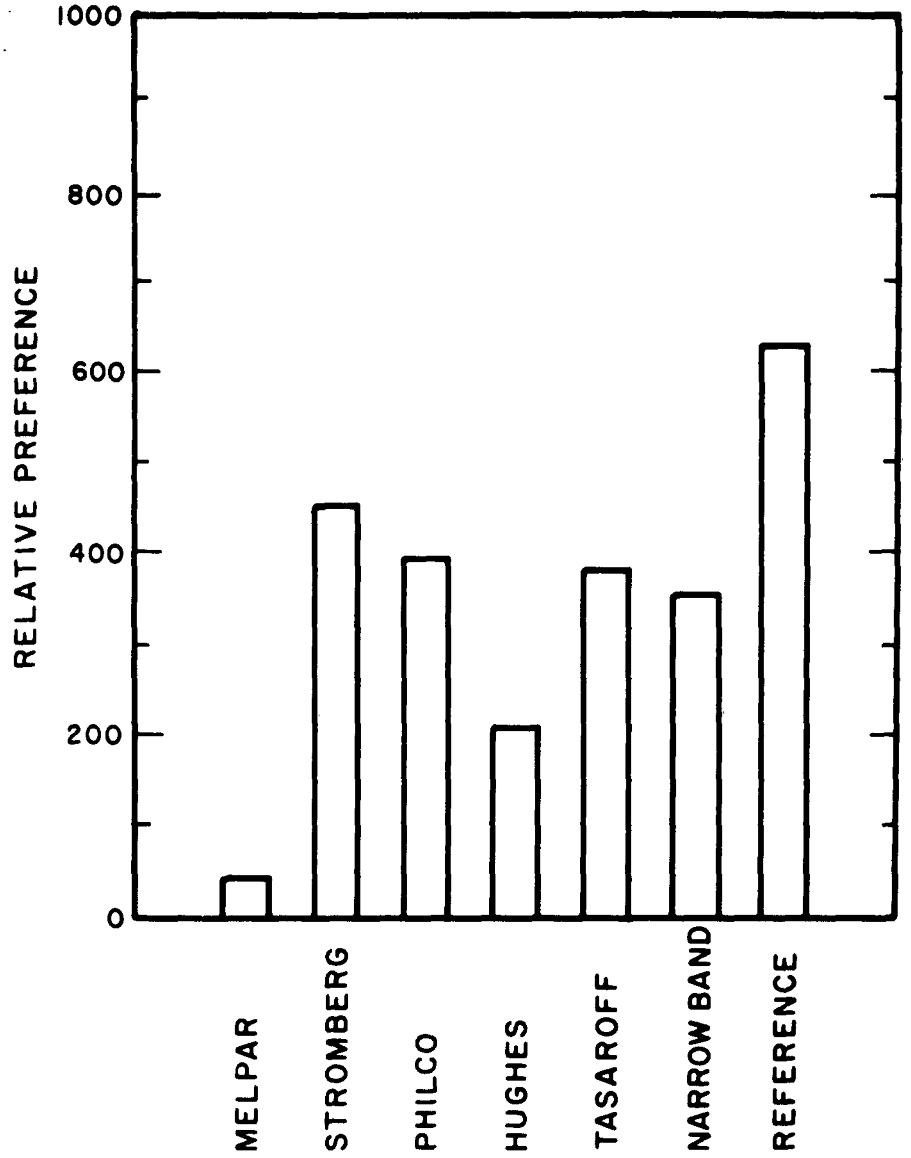


FIG. 2.5-2 RESULTS OF QUALITY TESTS WITH SAME SENTENCE (PB CARRIER PHRASE)

2.6 Talker Identification Tests

One particular measurement of the quality of the output of a system involves the ability of listeners to recognize specific talkers. A high quality system will leave intact those features of the speech wave which carry information about the talker, such as pitch variations, articulatory characteristics, and stress patterns.

The present talker identification tests may be conveniently divided into two groups. The first group of 14 tests was recorded with two quartets of speakers, so that each quartet could be tested over each of the seven systems. All tests commence with each of the four talkers identifying himself by number and reading two training sentences. The test proper consists of 20 randomized test sentences that constitute 20 items. Following a suitable pause after each item, during which the subjects record their responses, the previous talker identifies himself. This format ensures that learning continues throughout the course of the test.

The listeners were provided with the following set of instructions:

"Before the test begins, you will hear four talkers read two training sentences each to familiarize you with their voices. Each talker will identify himself by a number. A test item will consist of a sentence read by one of the four talkers. During the pause which follows each sentence you are to write the number of the talker you believe spoke next to the appropriate item on your answer sheet. At the end of the pause the actual talker will identify himself. If your answer was correct, make a check to the right of it. If your answer was incorrect, do not make any mark but wait for the next item. There are 20 items in the test."

This group of tests was administered to a crew of 30 listeners at Tufts University.

The second group of eight talker identification tests (four tests were recorded by each of two quartets) was administered to only 15 listeners. These tests were master recorded at Melpar, Inc. using their microphone facilities. While recordings were being made from the input to the Melpar system, the system output was simultaneously recorded for two tests. Later, two further tests that were recorded from the input to the Melpar system were played back through the same system, and the output was again recorded. Dubbings of four additional tests, originally recorded from the input to the Melpar system, were also played back through the Hughes and Philco channel vocoders. The tests in this second group have the same structure as those in the former group, except that each of the four talkers read five (instead of two) training sentences before the test items began.

One major difference between the talker identification tests recorded at our laboratories and those recorded at Melpar, Inc. is that for our recordings the microphone was suspended in the center of a circle of four talkers, about 30 inches from each talker's lips, whereas the microphone in the Melpar recordings was passed among the talkers and held 1 to 2 inches from the lips.

Although the tests are of the self-scoring type, most subjects were unable to correct their tests properly as they took them. Frequently the subjects did not give themselves credit where it was due and about equally often they gave themselves credit where their response was incorrect. This situation may be explained in terms of the fact that the correct answer, having been processed by the system under test, is not always clearly intelligible to the listener.

The mean scores (correct identification in percent) for the two groups of tests are given in Tables 2.6-1 and 2.6-2. These scores were obtained by correcting the subjects' "self-scored" answer sheets with the aid of appropriate lists that were used during the recording sessions.

Table 2.6-1
Mean Scores (Correct Talker Identification in Percent)
for 30 Subjects
Tests Master Recorded at BBN

System	Quartet I	Quartet II	Average for Both Quartets
Melpar	27	37	32
Stromberg	41	54	48
Philco	43	35	39
Hughes	33	31	32
Tasaroff-Daguet	45	44*	45
Narrow-Band	45	48	47
Reference	54	59	57

*Quartet III

Table 2.6-2
Mean Scores (Correct Talker Identification in Percent)
for 15 Subjects
Tests Master Recorded at Melpar, Inc.

System and Condition	Quartet IV	Quartet V	Average for Both Quartets
Melpar(live)	43	48	46
Melpar	31	31	31
Philco	47	55	51
Hughes	39	41	40

An analysis of variance of test score distributions was undertaken for the first group of tests (see Tables 5.2-5 and 5.2-7). The data obtained for each quartet were then examined to determine what mean score differences are statistically significant (see Tables 5.2-6 and 5.2-8). The order of some systems that fell into groups with insignificant differences between mean scores has been modified for both quartets in order to arrive at a rank order which has general validity. This rank order is shown in Table 2.6-3, together with the significant differences between system scores for quartets I and II.

Table 2.6-3
Rank Order of Systems, from Best to Worst,
According to Significant Differences between Scores
Obtained for Quartets I and II. Tests Master Recorded at BBN

Significant* Difference Quartet I	System	Significant* Difference Quartet II
	Reference	
Yes		No
	Stromberg	
No		Yes
	Narrow-Band	
No		No
	Tasaroff-Daguet	
No		Yes
	Philco	
Yes		No
	Hughes	
Yes		No
	Melpar	

(*Statistically significant at the $p \leq 0.05$ level of confidence.)

The results obtained for the second group of tests, master recorded at Melpar, Inc., indicate that the use of the microphone for which the Melpar system was designed does not improve the low rating of this system. The scores for the Hughes and Philco systems are, however, improved somewhat by using the Melpar microphone facilities. Talker identification via the Melpar system is significantly improved when the system is tested live, i.e., when the recording microphone is substituted for a pre-recorded tape.

2.7 Comprehension of Continuous Speech in Noise

The context in which an unintelligible word occurs is often important because it may help the listener to resolve his doubts about the ambiguous word. A test in which listeners are asked whether they can make out the "essense" of a message may therefore result in much higher scores than a test in which the intelligibility of isolated words is measured. Tests dealing with the comprehension of continuous speech are different from intelligibility and quality tests; they must be considered in a separate category.

The relative comprehension of continuous speech was measured for various signal-to-noise ratio conditions at the inputs of the seven speech compression systems considered. One test consisting of 49 samples of continuous speech, each set against a constant background of noise, was administered to a crew of 35 listeners. Each test sample had a duration of 25 seconds and a signal-to-noise ratio that was either 30, 25, 20, 15, 10, 5 or 0 db. The subjects were instructed to mark the items on their answer sheets with an A if they could "make out almost every word," with a B if they could "make out only a few words," and with a C if they could "make out almost no word at all." The speech material, which was recorded by a male talker under quiet conditions, was selected from: A. Smith, The Wealth of Nations.* An Altec-Lansing Model 661A dynamic microphone was positioned 10 inches from the talker's lips. The noise had a spectrum similar to that of the long-time average of speech and was electrically added to the recorded speech signal.

The results for this test are shown in Table 2.7-1, and in graphical form in Fig. 2.7-1. For a given system and a particular signal-to-noise ratio, the score A was weighted 2 points, the score B, 1 point, and the score C, zero. The ordinate scale in Fig. 2.7-1 has been arbitrarily selected so that the reference system scores 100% (relative comprehension) for a signal-to-noise ratio of 30 db.

Although the results vary somewhat at different signal-to-noise ratios, it appears that the Hughes, Philco, Tasaroff-Daguet and Narrow-Band systems perform, on the average, about equally well on this test. The Reference system and the Stromberg semi-vocoder are much less adversely affected by noise than are the other systems, and the Melpar formant vocoder performs very poorly in noise. One possible explanation for the relatively poor performance of the formant vocoder is that the talker reads at a rate that is too high for proper operation of the pitch extractor and formant-tracking circuits in the device. The Melpar vocoder apparently has considerable difficulty in tracking the fundamental and formant frequencies for speech spoken at a fast rate. It will be recalled that for PB word tests, where each test word is spoken individually, intelligibility was still appreciable in a 15 db signal-to-noise ratio condition, whereas relative comprehension of this continuous discourse is essentially zero under similar noise conditions.

* E. P. Dutton and Co., Inc., New York, 1937

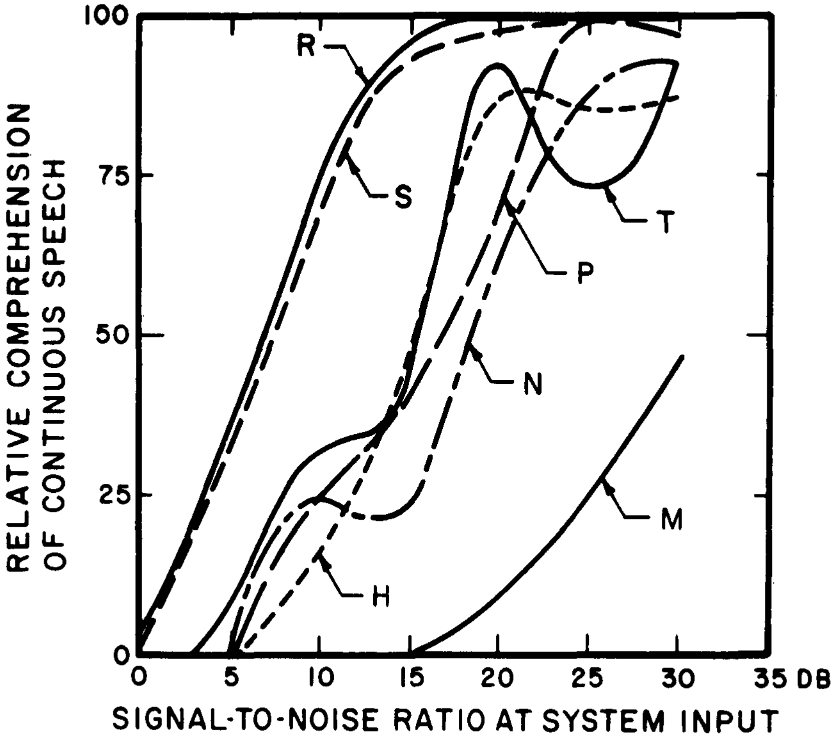
Table 2.7-1

Results of Relative Comprehension Test

Each Entry Represents a Cumulative Point Score for 35 Subjects

System	Signal-to-Noise Ratio (db)						
	30	25	20	15	10	5	0
Melpar	33	18	7	0	0	0	0
Stromberg	70	69	67	68	41	24	2
Philco	68	70	48	29	18	0	0
Hughes	62	60	63	32	11	0	0
Tasaroff-Daguet	65	52	65	31	23	5	1
Narrow-Band	65	61	42	14	19	1	1
Reference	70*	70	70	66	55	23	3

*All 35 subjects rated this sample "A" (2 points).



KEY

M - MELPAR

S - STROMBERG

P - PHILCO

H - HUGHES

T - TASAROFF

N - NARROW BAND

R - REFERENCE

FIG. 2.7-1 FINAL RESULTS OF RELATIVE COMPREHENSION TEST, BASED ON DATA GIVEN IN TABLE 2.7-1

3. EVALUATION OF SPEECH COMPRESSION SYSTEMS TESTED

3.1 Reference (Low-Pass) System

The reference system consists of a low-pass filter with a cut-off frequency of 1500 cps. The slope of the filter characteristic above this frequency is 36 db/octave, and thus the gain is down about 18 db at 2100 cps and 36 db at 3000 cps.

The types of errors in identification of vowels and consonants in nonsense syllables for the low-pass system are those that would be expected in view of the lack of high-frequency data in the speech signal. There are a few errors in voicing and manner of production, especially for fricative and stop consonants since the high-frequency energy for such consonants apparently has some cue value for these distinctions. Errors in manner of production for nasals, liquids and glides are relatively small, since these voiced sounds have little high-frequency energy.

The identification of place of consonant production for the low-pass system is relatively good for consonants with appreciable low-frequency energy, namely the nasals, liquids and glides. For voiceless fricatives, on the other hand, the number of errors is large (22 percent for /fsʃ/, since distinctions among these consonants are apparently made primarily on the basis of high-frequency spectrum shape. The errors in place for voiceless stops and for voiced fricatives and stops are in the range 12-15 percent, i.e., somewhat less than for voiceless fricatives; cues for these consonants are carried in part by low frequencies, including the vowel transitions, and in part by high frequencies. As for the glides, /j/ is frequently identified as /w/, since the high-frequency energy concentration in /j/ is not passed by the filter.

For vowels, the number of errors is considerably smaller than for any of the other systems. Such a result is to be expected, since the important cues for vowels -- the frequency locations for the first two formants -- are known to lie below 2500 cps.

In general, most of the error scores for nonsense syllables processed by the reference system were comparable to or lower than those for all other systems. The scores for this system were significantly poorer than those for other systems only for voiceless fricatives. The tests other than nonsense syllables also demonstrate the superior performance of the reference system relative to all other systems. This superiority seems especially clear-cut for the voice quality tests (Figs. 2.5-1 and 2.5-2) and for the talker identification tests summarized in Table 2.6-3. (The scores for PB tests with the reference system do not greatly exceed those for some of the other systems if PB scores for the reference system are adjusted downward by 5 to 10 percentage points to correct for learning effects.) Such a result is reasonable if it is hypothesized that voice quality and ability to identify talkers are preserved if the temporal properties of the signal, particularly the detailed properties of the quasi-periodic voice source, are not distorted. The validity of this hypothesis will be examined further as data from other systems are discussed.

3.2 Channel Vocoder

The vocoder technique is based on a model that views the generation of speech as the excitation of the vocal cavities by either a noise source or a quasi-periodic buzz source.[8,9] This generation process is simulated in the channel vocoder by using electrical versions of one or another of such sources as excitation for a bank of filters and by varying the gain of each filter to provide an appropriate spectral output as a function of time. If a maximum amount of bandwidth compression is to be achieved, the number of filters must not be too large, and the rate of change of the signals that specify the gains of the individual filters must be limited. Thus, any vocoder design must represent a compromise between the quality and intelligibility of the speech output on one hand and the channel capacity on the other.

The bandwidth and information rate required for a channel vocoder have been estimated previously.[6] The bandwidth of the transmitted signal corresponding to each filter channel is considered to be about 25 cps, and the bandwidth required for the signal that indicates the fundamental frequency is apparently also about 25 cps. For a 16-channel vocoder, therefore, the total bandwidth is about 425 cps. In the case of a digitized vocoder, sampling rates of about 45 samples/sec for each channel, with 3-bit amplitude specification, seem adequate for the channel signals, whereas 6 bits are necessary for the "pitch" signal. Thus the total information rate is about 2400 bits/sec.

The following discussion of the channel vocoder performance is based on the results obtained for the Philco 16-channel vocoder. The filters in this vocoder are uniformly spaced up to 1000 cps, with bandwidths of 133 cps, and are logarithmically spaced above 1000 cps to an upper frequency limit of 3800 cps. [It was evident in the course of the testing program that the 12-channel digitized Hughes Vocoder had a malfunction, and thus the results for that system were not considered to be representative of the performance of channel vocoders in general. Several aspects of the test data provide evidence that the Hughes Vocoder was not operating in an optimum way. For example, the average PB word score in quiet for two male talkers was only 61 percent (Fig. 2.3-2), whereas for a signal-to-noise ratio of 15 db the average score (for one talker) was much higher -- 73 percent (Fig. 2.3-4). Informal tests of the Hughes system for analog operation showed greatly improved performance compared with digital operation for talkers in quiet -- of the order of 15 to 20 percent improvement for PB words. Apparently, therefore, the malfunction was in the digital circuits.]

Some inherent limitations in present channel vocoder techniques are evident from the data for the nonsense syllable tests. Consider first the voiced-voiceless distinction, for which the probability of error is 3.6 percent for the Philco Channel Vocoder. It is suggested that the necessity of making this binary decision in the vocoder analyzer introduces an error that is difficult to reduce much below this value without appreciable increase in complexity. Presumably part of the voiced-voiceless error is due to the limitation that buzz and noise source cannot exist simultaneously in the vocoder, whereas both sources may, of course, exist for certain speech sounds. Part of the error is probably also attributable to the difficulty of devising a pitch extractor that operates reliably and with negligible delay. A delay of as little as 10 to 20 msec in operation of the buzz source could, for example, easily result in a voiced fricative or stop being called voiceless. In contrast to the conventional channel vocoder are the semi-vocoder, the reference system, the Tasaroff-Daguet system, and the spectrum sampling system, none of which require a buzz-hiss decision or pitch extraction. These systems all have fewer errors in the voiced-voiceless distinction than does the channel vocoder.

The feature interrupted-continuant is detected incorrectly about 8.9 percent of the time for the channel vocoder. Errors in this feature may arise because the modulators controlling the individual filter outputs cannot change rapidly due to limitations in bandwidth and sampling rate of the transmission signals. Note, for example, that voiced stops are frequently called voiced fricatives, since the rapid changes inherent in the stops cannot, apparently, be reproduced in the vocoder. The limited dynamic range of the system apparently also contributes to errors in the interrupted-continuant feature; weak sounds such as /f/ are often not reproduced at all, with the result that they are heard as stops. Arguments similar to these could be given to indicate probable causes of error in judging manner of production between nasals, liquids and glides (about 8 percent probability of error). To some extent the grossness of the reproduced frequency spectrum, as well as inadequacies in reproduction of temporal variations of the spectrum, could contribute to errors in this decision.

Errors in judging place of production of many of the consonants and vowels processed by the channel vocoder are attributable largely to the fact that the acoustic spectrum is analyzed and reproduced only grossly by the filter banks. For voiceless and voiced fricatives, these errors seem relatively small (5.9 percent and 3.2 percent for these tests), indicating that the gross spectral features as reproduced by the channel vocoder are adequate for these classes of sounds.

In cases where important cues are contributed by rapidly changing spectral features, however, the performance of the channel vocoder is less satisfactory. Examples are the /f/ - /θ/ distinction (37 percent errors) which seems to depend on formant transitions, the voiceless stops (14 percent), the voiced stops (17 percent), and the nasals (18 percent errors for initial and 40 percent for final). For these groups of sounds, both the grossness of the spectral representation and inadequate reproduction of rapid temporal changes contribute to the errors.

As for the vowels, relatively few errors are made in the tests involving the front vowel sequence /i ɪ ɛ æ / and the back vowel sequence /u ʊ ʌ ɑ/. Vowels within these groups differ in first formant frequency, and hence have quite different over-all spectrum shapes. The spectra are apparently reproduced with sufficient accuracy to permit discrimination among these vowels. For the long vowels /i æ ɑ u/ and short vowels , ɪ ɛ ʌ ʊ /, however, the number of errors is much greater (14 and 15 percent, respectively). Most of the errors are made within pairs of vowels that have roughly the same first formant frequencies and differ only in the second formant frequencies. Examples of such pairs are /æ ɑ/, /ɛ ʌ/ and /ɪ ʊ/.

For the Philco vocoder, the PB word intelligibility in quiet for male talkers was 85 percent, and thus was comparable to that of the semi-vocoder. As noted above, there is some loss in intelligibility in the channel vocoder relative to the semi-vocoder due to errors in reproducing excitation characteristics. However, the filter channels for the Philco channel vocoder extend to a higher frequency than those for the Stromberg semi-vocoder, and hence the errors for certain sounds with high-frequency energy are greater for this semi-vocoder than for the channel vocoder. Furthermore, the method of extracting an excitation signal in the semi-vocoder may not yield a sufficiently broad-band signal, and thus there may be a lack of high frequencies in the speech synthesized at the receiver. The result is that the overall intelligibilities for the two systems are comparable.

The PB scores for female talkers are much lower than those for male talkers (59 versus 85 percent). This deterioration is probably attributable, at least in part, to poor performance of the pitch extractor for female voices for which the fundamental frequency is high.

In the quality tests and in the talker identification tests, the performance of the semi-vocoder was significantly superior to that of the channel vocoder, as shown in Figs. 2.5-1, 2.5-2, and Table 2.6-3. These data provide further evidence, therefore, that good voice quality and correct talker identification depend on maintaining an accurate replica of the temporal properties of the voice excitation. For any system in which the fundamental frequency must be extracted and the source must be reconstituted at the receiver, the voice quality deteriorates.

3.3 Semi-Vocoder

The principle of operation of the semi-vocoder is similar to that of the channel vocoder, with the exception that no pitch extractor or voice-hiss detector is required in the former.[47] In the particular version tested, a baseband covering the frequency range 250 to 750 cps is transmitted directly, and by various distortion means at the receiver this signal is used to form a relatively flat-spectrum excitation signal for a set of 13 conventional vocoder channels in the frequency range 750 to 3250 cps. The baseband is also reproduced directly at the receiving end of the link and is mixed with the signal synthesized by the filters.

The analog semi-vocoder that was tested required a bandwidth of about 900 cps for transmission of both baseband and vocoded signals, including guard bands. (The system was actually designed in such a way that three sets of semi-vocoder signals, suitably multiplexed, could be transmitted via a conventional telephone link.) The information rate required for a digitized version of such a semi-vocoder has been estimated to be about 6500 bits per sec.

As noted in the discussion of the channel vocoder, the number of errors in the nonsense syllable tests for the *voiced-voiceless* distinction (1.4 percent) is much smaller for the semi-vocoder than for the conventional channel vocoder. The buzz-hiss distinction is made automatically in the semi-vocoder analyzer, and no decision process is required in the equipment. Likewise, the features involving *manner* of production are identified with a somewhat better score for the semi-vocoder than for the channel vocoder, indicating the improvement associated with direct transmission of the baseband.

With regard to identification of *place* of production, the errors for voiceless consonants are relatively high since the upper frequency range for the semi-vocoder tested was only 3250 cps. Also, as noted previously, there may be some difficulty in deriving from the 250-750 cps baseband a suitable noise-excitation signal with enough high-frequency energy. Direct transmission of the baseband helps to reduce the errors in identifying place of production of voiced sounds, particularly the nasals, glides, liquids and vowels. The greatest number of errors for *vowels* occurs for cases where two vowels have roughly the same first-formant frequencies but have minimal differences in second-formant frequencies, such as the pairs /i-u/, /æ - ɑ/, /ɪ-ʊ/, and /ɛ-ʌ/.

The baseband in the semi-vocoder tested, i.e., 250-750 cps, was apparently selected on the assumption that the speech input was restricted at low frequencies by a carbon microphone and that the equipment would be used by male talkers. For the high fundamental frequency used by female talkers, the baseband might contain two and sometimes only one harmonic of the fundamental. Under such circumstances, it becomes difficult to devise a distorting circuit that yields an excitation signal with a flat spectrum envelope, indicating equal amplitude for all harmonics of the fundamental. The sharp drop in intelligibility of the semi-vocoder output for female talkers relative to male talkers (56 and 86 percent, respectively) reflects this limitation.

When speech was mixed with noise at the semi-vocoder input, the PB word intelligibility for a 15 db signal-to-noise ratio decreased from 86 to 70 percent for male voices, while the "relative comprehension" of continuous speech showed only a small decrease for the same noise conditions. It appears that the features upon which the listener bases his judgments in the comprehension test include, to a large extent, the stress and intonation pattern. These patterns are preserved reasonably well in the case of the semi-vocoder, even though some other features may be partially obscured by the noise. The results of the quality and speaker-identification tests also reflect the degree to which these patterns are preserved in the semi-vocoder. The semi-vocoder ranks highest in both types of tests when compared to the other compression systems.

3.4 Formant Vocoder

The formant vocoder represents a modification of the basic channel vocoder. As in the channel vocoder, a circuit in the analyzer makes the distinction between buzz and hiss excitation; a signal proportional to the fundamental frequency is extracted for intervals in which there is buzz excitation, and this signal is transmitted to the receiver. Spectral information is described by a relatively small number of parameters that indicate certain salient spectral features. During non-nasal vowel or vowel-like sounds, two or three of these signals are supposed to indicate the frequencies of the lowest two or three vocal-tract resonances or formants. In some versions of formant vocoders, the amplitudes of the resonances are specified as well as their frequencies. When vowels are nasalized, and at times when the vocal-tract excitation is not at the glottis, the way in which spectral information is extracted and synthesized is somewhat different from one version of the formant vocoder to another.

For the formant vocoder tested in this program, seven parameters were extracted at the analyzer: amplitudes and frequencies of formants 1 and 2, location of a high-frequency "fricative formant," amplitude of high-frequency portion of the signal, and fundamental frequency. The analog bandwidth required for each of these channels was estimated to be 20 cps, giving a total analog bandwidth of 140 cps. For digital operation, the sampling rate was 43.5 cps, and all parameters except fundamental frequency were quantized to 3 bits; 5 bits were used to code fundamental frequency. Thus, the total information rate was 1000 bits/sec.

While the details of the analysis and synthesis procedures may vary greatly from one type of formant vocoder to another, the data obtained for the particular version tested in the present program seem to illustrate some of the limitations inherent in formant vocoders in general.

We shall examine first the performance for non-nasal vowels and for the vowel-like sounds /w j ℓ r/, since the formant vocoder technique is designed to reproduce such sounds in a reasonably straightforward way. The group of sounds for which the fewest errors are made is the back vowel series /ɑ ʌ ʊ u/. These vowels can be said to be distinguished on the basis of the frequency position of a main concentration of energy in the frequency range 200 to 1200 cps. Apparently the formant trackers detect this energy concentration adequately, and a correct identification is made within this group of vowels whether one or two formants are assigned to this region.

A large number of errors are made, however, for the two vowel groups /i æ ɑ u/ and /ɪ ɛ ʌ ʊ/caused, apparently, by errors in tracking the second formant. In the case of /i/, for example, two formants seem to be often assigned to the strong energy concentration at low frequencies. For other vowels there is a tendency for the first two formants to be called one formant when they are closely spaced, such that the second formant is then assigned incorrectly. The identification of the liquids /ℓ r/is quite poor, since the particular system tested was inherently unable to track or to generate the low-frequency third formant that is characteristic of /r/. The glides /w j/ are extreme examples of cases where there are two closely-spaced formants (the first two for /w/, the second and third for /j/, and it is clear from the results for these sounds that the formants are not tracked correctly.

When important cues for identification come from rapid formant transitions, as in consonant classes such as voiced stops and nasals, the percentage of errors is higher than for all other classes of consonants. It would appear that substantial tracking errors are made when changes in formant frequency occur in a few tens of milliseconds. For example, 61 percent errors are made in identifying place of production of /b d g/. Of this group, the response that is made more often than the others is /g/, which is characterized by formants that move less rapidly than those for /b/ and /d/. Likewise, the nasals are identified only slightly above chance level. Errors in place of production for voiceless fricatives (13 percent for /f s ʃ/)are not as high as for other classes of consonants, since a special circuit was incorporated in the synthesizer to accommodate this class of sounds. Voiceless stops (49 percent error) and voiced fricatives (32 percent error), however, are not identified as accurately as voiceless fricatives.

The data obtained from the PB word intelligibility tests, the quality judgments and the talker-identification tests all indicate that the performance of the formant vocoder was poorer than that of all other systems tested. This performance is the result of errors both in reproduction of fundamental frequency and in reproduction of the spectrum. Furthermore, the tests on the comprehension of continuous speech in the presence of noise indicate that noise mixed with the input speech results in a sharp increase in the errors in tracking the various parameters. Apparently the method used to track formants in the system tested (a method based on measurement of average density of zero-crossings in particular frequency bands) was rather sensitive to noise at the input.

3.5 Spectrum Sampling (Narrow Band) System

In the spectrum sampling system the speech is passed through several narrow frequency bands distributed throughout the speech frequency range, and the resulting signal is transmitted to the receiver.[27] The hypothesis is that, if the sampled frequency bands are properly selected, the listener may perform some sort of spectral interpolation such that the intelligibility is much greater than that of a single continuous frequency band of the same total width. In the particular system studied in the present series of tests, three frequency bands were used as follows: 400-800, 1550-1950 and 3425-3550 cps at points 30 db down from the mid-band gain, giving a total analog bandwidth of 925 cps by this measure. However, in the transposition of these three bands into a single continuous band for transmission, the bands were overlapped slightly, so that a total continuous bandwidth of only about 800 cps was used. Other combinations of bands, possibly six bands instead of three, could have been selected to give greater or less total bandwidth and different overall intelligibility. At the time the present tests were conducted a system of only three bands was available.

The results of the nonsense syllable tests indicate that, in comparison with other systems with comparable bandwidth, the band selection system reproduces the voiced-voiceless distinction and the manner of articulation reasonably well (2.4 percent errors for voiced-voiceless, 4.3 percent for interrupted-continuant). As would be expected, any distinction that depends primarily on temporal rather than detailed spectral characteristics of the signal should be received with a fairly small number of errors for this system, since it imposes no distortion on the temporal characteristics.

The number of errors in identification of place of production for vowels and consonants is, however, quite high in comparison with most of the other systems, as Fig. 2.4-3 snows. The types of errors seem always to follow a pattern that is closely related to the particular frequency bands used in the system. For example, relatively few errors are made in identifying /s/ and /ʃ/, whereas /f/ is identified as /s/ most of the time. The fricative /ʃ/ has a major concentration of energy around 2000 cps, and apparently this is adequately reproduced by the 1550-1950 cps frequency band. A spectral energy maximum around 3500 cps could, on the other hand, lead to an acceptable /s/. For /f/, the spectrum is usually rather flat in the frequency range up to 5000 cps. Introduction of an artificial peak around 3500 cps could easily lead to erroneous identification of /f/ and /s/. A similar pattern of response is found for the voiced fricatives /v z ʒ/.

The errors in identification of stop and nasal consonants are patterned in a way that indicated a high accuracy of identification of the post-dental consonants /t d n/, with many more errors for the bilabials and velars. Cues for the identification of these consonants are known to be carried by formant transitions of the adjacent vowel, particularly transitions of the second formant. For post-dental consonats, the locus or target frequency of the second formant is known to be in the vicinity of 1800 cps. The small number of errors for post-dental consonants apparently arises, therefore, from the fact that this locus frequency is in one of the frequency bands passed by the narrow band system. Performance for the other consonants in stop and nasal classes is, however, quite poor, with the result that the overall errors in place of production for these consonants are in the range 30 to 40 percent.

The pattern of errors for <u>vowels</u> is likewise explainable in terms of the frequency ranges of the lower two filters in relation to the frequencies of the first two vowel formants. In general it can be said that the formant frequencies for the front vowels /i ɪ ɛ æ/ lie within the frequency ranges of the lower two filters (/i/ is slightly outside the range), and the overall error score for these vowels in all tests was about 8 percent. On the other hand, the formant frequencies of the second formants for the back vowels are always below the frequency range 1550-1950 cps passed by the system. This is reflected in the overall error score for the back vowels in all vowel tests, which was about 35 percent.

4. STATUS OF VARIOUS SPEECH COMPRESSION TECHNIQUES AND RECOMMENDATIONS FOR FUTURE RESEARCH AND DEVELOPMENT

4.1 Introduction

To facilitate a discussion of the merits and possible future potential of presently available speech compression techniques, it is convenient to group the techniques according to the degree of compression they achieve. Five groups have been arbitrarily defined, as shown in Table 4.1-1. Each group includes those techniques which require an information rate in the indicated range for digital operation, or a corresponding bandwidth in the indicated range for analog operation.* The techniques under each group heading will be discussed individually and evaluated with respect to their relative strength, their possible future potential, and their readiness for equipment development. In addition, recommendations for more research effort on specific systems will be made wherever applicable. Finally, Section 4.7 will give a survey of current research that is relevent to the development of speech compression systems in general.

The relative strength of a particular technique is estimated partly on the basis of test results obtained from representative compression systems. Particular attention was given to results from PB word intelligibility tests and voice quality tests. The relative complexity of the technique is also considered in this

* The relation between bandwidth for analog operation and information rate for digital operation is not invariant, and depends upon the manner in which the analog signal is coded. For many of the compression systems, the analog signals are sampled periodically at the Nyquist rate of about two times the analog bandwidth, and the amplitudes of the quantized samples are specified by three to five bits. This rule was used to determine the corresponding ranges of bandwidth and information rate in Table 4.1-1.

estimate since the amount of equipment associated with a given technique may restrict its practical value and application. The necessary data on PB word intelligibility, voice quality, and complexity were obtained largely from the present studies, particularly for the systems actually tested, but complementary information was also obtained from studies that have been reported previously and from other sources.[13,36,51]

The discussion to be given in the following sections will indicate that some speech compression techniques are now ready for equipment development, in particular the semi-vocoder, the spectrum sampling scheme, and the channel vocoder, while other techniques are in need of more research. Research on the semi-vocoder and channel vocoder techniques should, however, continue with a view to obtaining further improvements in performance. Those techniques providing an intermediate degree of compression (Groups B and C) are, in general, the most promising for the immediate future, although techniques offering more compression will be improved and may become more attractive at a later time. On the immediate horizon is the spectrum matching and coding procedure, which may, through the use of rather complex terminal equipment, achieve appreciable compression with performance comparable to that of the channel vocoder. Efforts for equipment development of particular high-compression systems such as the formant vocoder are still somewhat premature, considering the number of unresolved theoretical questions associated with these relatively new approaches.

Table 4.1-1.

Presently Available Speech Compression Techniques Grouped According to Degree of Compression Achieved

Group	A	B	C	D	E
Range of Information Rate (bits/sec)	18,000 to 12,000	12,000 to 5,000	5,000 to 2,000	2,000 to 800	Below 800
Approximate Range of Analog Bandwidth (cps)	2,000 to 1,500	1,500 to 600	600 to 250	250 to 100	Under 100
Techniques	Band-Pass Filtering Amplitude Clipping Extremal Coding Time Compression Semi-Vocoder Spectrum Sampling	Semi-Vocoder Spectrum Sampling Tasaroff-Daguet	Channel Vocoder Cross- and Auto-correlation Vocoders	Formant Vocoder "Peak-Picker"	Spectrum Pattern Matching and Coding

4.2 Group A: 18,000 - 12,000 bits/sec (2,000 - 1,500 cps)

Band-pass filtering. This is perhaps the simplest approach to the problem of speech compression, and one which has been studied in great detail. Experiments with an optimally-centered 2,000 cps wide band-pass filter suggest that PB-word intelligibility scores of 85-90% are readily obtainable, although the voice quality is slightly inferior to that of a conventional telephone channel. Because of the extremely low compression that can be achieved by band-pass filtering, the commercial usefulness of this technique is obviously very limited. The technique is of some value, however, for comparison purposes. The efficiency of another compression technique may be evaluated in terms of the bandwidth it requires when compared to the bandwidth of an equally intelligible system consisting of a single, optimally-centered band-pass filter.

Amplitude clipping. Amplitude clipping refers to a process whereby the input speech waveform is amplified linearly up to a specified amplitude level; beyond this level the output signal does not increase with further increases in the input signal. A modest amount of amplitude clipping has little effect on the intelligibility of speech. The technique is employed mainly to extend the effective range of radio-telephone transmitters. Infinite clipping, which reduces the speech signal to a rectangular waveform, gives a PB word intelligibility near 80% and a rather unpleasant, harsh voice quality. Differentiation of the speech signal before clipping improves the intelligibility, especially in the presence of noise, and integration after clipping improves the quality somewhat, but integration before clipping lowers the intelligibility drastically. Licklider[31] has determined that amplitude-dichotomized, time-quantized speech waves are reasonably intelligible for information rates above 8,000 - 10,000 bits per sec. Compression systems based on amplitude clipping could therefore be categorized either in Group A or Group B.

Extremal coding.[33] This is a digital scheme for speech transmission that is related to time-quantized clipped speech. The primary advantages of this technique are a telephone-like voice quality and a relatively high PB word intelligibility -- approaching 90%. The technique is complex, however, and because no working model of a system has been constructed, all operations have been simulated on a digital computer. The extreme amplitudes of the speech wave and the time intervals between these extremes must be extracted, and a buffer memory is required to convert the randomly occurring information about the extremes to a uniform rate before transmission. There exists a possible future potential for this technique in communication links with large terminal installations having computing and PCM facilities. Equipment development may commence in this area without further research effort.

Time compression. This procedure involves the periodic extraction of time samples from the speech signal. These samples are divided in frequency, abutted in time, and stored for later reproduction at a speed appropriate for restoration of the speech. With a properly chosen sampling period, a moderate amount of time compression can be achieved with a PB-word intelligibility somewhat better than 80%. The voice quality is probably superior to that of clipped speech, although this technique is inherently more complex. Various schemes of time compression have been investigated in detail, and it appears that the commercial possibilities for the technique in obtaining a bandwidth compression of a factor of two or more are very small. Further equipment development and research are therefore not recommended.

Semi-vocoder (base-band or voice-excited vocoder). The semi-vocoder represents a compression technique which clearly has possible future potential. PB-word scores near 90% and a good, telephone-like voice quality are the main advantages of this technique. Semi-vocoders can be designed to use available communication channels having bandwidths anywhere in the range 800-3000 cps, depending on the intelligibility and voice quality required. For example, "hi-fi" voice transmission has been achieved with a semi-vocoder that uses a conventional telephone channel.[47] The filter banks required at the analysis and synthesis terminals of the semi-vocoder constitute a limitation on this technique, since it is difficult to build light and compact terminal equipment when a number of filters must be included. Although the principles of the semi-vocoder are well established, it is suggested that further research could profitably be carried out in order to arrive at optimum filter arrangements, optimum filter characteristics, and optimum procedures for deriving an excitation signal from the baseband. Thus, for a given total bandwidth for the transmission signals and for given channel characteristics, each of these features could be adjusted to maximize the intelligibility and voice quality. Since the modification of any one feature of the equipment would probably change the intelligibility of only a small number of speech sounds, short nonsense syllable intelligibility tests of the type described in Section 2.4 could be used during this research phase. Some of this work has already been carried out or is in progress, but it is suggested that further studies could lead to improved semi-vocoder performance. At the same time, it is clear that the semi-vocoder technique is sufficiently well advanced that equipment development may be scheduled concurrent with the research.

Spectrum sampling. This is a relatively simple technique involving band-pass filtering of two or more radio-frequency carriers which are amplitude-modulated by the speech signal. The filter outputs are demodulated and summed to reproduce selected regions of the original speech spectrum. With a system having three 650 cps wide filters centered about optimum frequencies (500, 1500, 2500 cps), PB word scores near 90% have been obtained. With an 8-filter system (1500 cps bandwidth*) PB word scores of 90-95% are possible. In general it may be said that the nominal bandwidth required for spectrum sampling with an optimum number and location of filters is approximately equal to one-half of the bandwidth of an optimally centered band-pass filter providing the same level of intelligibility. The spectrum sampling technique has potential, therefore, in applications where only a modest amount of compression is required, and where simple, compact and light terminal equipment is a necessity. It cannot compete with the semi-vocoder, however, in situations where more complex terminal equipment can be allowed.

4.3 Group B: 12,000 - 5,000 bits/sec (1,500 - 600 cps)

The Semi-vocoder. The semi-vocoder is mentioned again in this group because it is essentially capable of more compression than is represented by a bandwidth of 1500 cps. The semi-vocoder that was tested required an analog bandwidth of 900 cps for transmission of both the base-band and the vocoder channel information, including guard bands.

Comments regarding the potential of the semi-vocoder and the research required to improve the semi-vocoder performance have been given in section 4.2.

Spectrum sampling. The spectrum sampling system which was tested uses three filters and has a nominal bandwidth of 800 cps. PB-word scores approaching 70% were obtained for male speakers, and the

* measured at the 30 db down points.

voice quality was inferior to that of the tested semi-vocoder. More extensive experiments have indicated that optimum performance (PB-word scores near 85%) is reached with seven or more filters for a nominal bandwidth of about 1000 cps. In the hope of being able to provide more compression without loss in intelligibility, a modification of the inherently simple spectrum sampling technique has been proposed. This modification involves moving the center frequencies of one or more of the sampling filters according to the short-time energy distribution in the speech spectrum. Some research needs to be done to determine the effectiveness of this operation. If a production model of the basic technique with fixed filters were contemplated, however, equipment development could commence without further research.

Tasaroff-Daguet. The Tasaroff-Daguet system is described and discussed in classified Sections 6.1 through 6.3.

4.4 Group C: 5,000 - 2,000 bits/sec (600 - 250 cps)

Channel vocoder. The sixteen-channel vocoder provides a PB-word intelligibility approaching 85% and a voice quality comparable to that obtained for the spectrum sampling system described in Group B. The voice quality of this channel vocoder is not as good as that of the semi-vocoder because of errors in pitch extraction at the sending terminal, errors in voiced-unvoiced switching at the synthesis terminal, and less accurate reproduction of low-frequency components of the signal. The channel vocoder is inherently more complex than the semi-vocoder, since, in addition to the filters, it requires equipment for coding the excitation signal.

It appears that the past and present research on this technique will probably lead to further improvements in performance. Several problems are now being investigated, and these and others require further research: (a) Studies are needed to determine what features of the excitation signal should be reproduced at the synthesizer in order to obtain maximum intelligibility and highest voice quality. (b) Based on information obtained in (a), procedures must be devised for extracting appropriate features of the excitation signal. (c) Further studies are needed to determine the arrangement of filters and filter characteristics that will give optimum performance for a given total channel bandwidth. Relative to (c), a recent review of speech compression techniques[13] makes the following statement:

> Recent developments have indicated that the steep, flat-bottomed band-pass filter characteristics commonly found in early vocoders are not necessary, and, on the contrary, that the quality of the synthesized speech may be improved by introducing simple, narrow-band tuned circuits. The band-pass filters at the analysis end require somewhat greater selectivity than those at the synthesizer. Other experiments have indicated that a dynamic expansion of channel signals, exaggerating differences in spectral levels, can improve the quality of conventional vocoders.

This research should be carried out with a view to determining, in a precise way, the types of distortion that are being imposed on the speech signal and the effect of these distortions on the perception of the various features that contribute to vowel and consonant intelligibility and to voice quality. It will then be possible to make modifications in the equipment to minimize the perceptual consequences of the distortions. Although research of this type is currently in progress and further research is proposed, it is suggested that the technique as it stands has sufficient merit to warrant equipment development. Such development has, of course, already been carried out to some extent, but as research on vocoders progresses the results of the research should be applied to the development of new equipment.

Cross- and Auto-correlation Vocoders. These systems are, in a sense, time-domain versions of the channel vocoder. In both schemes the fundamental frequency must be extracted and the voiced-voiceless distinction must be made. The schemes differ from the channel vocoder in that the transmitted signals specify the waveform in each period of the fundamental rather than the gross spectrum. With further research it is possible that the performance of the cross- and auto-correlation vocoders can reach, but probably not exceed, the maximum performance of conventional channel vocoders. An important advantage of these time-domain schemes over the channel vocoder is that banks of filters are not required either in the analyzer or in the synthesizer, and hence the equipment is smaller and less complex.

4.5 Group D: 2,000 - 800 bits/sec (250 - 100 cps)

Formant Vocoder. This vocoder represents a technique which is considerably more involved and not yet as refined as the technique of the conventional channel vocoder. The performance of the formant vocoder from the point of view of both intelligibility and voice quality is below that which would be considered acceptable for any but the most restricted applications. Further research on the nature of speech and further progress in techniques of speech analysis and synthesis are required before development of a complete system suitable for practical use is commenced. The experiments performed in connection with the present study, as well as other studies reported in the literature, indicate several areas where research is needed. Again, some of this research is already in progress, but more long-range research effort is needed before a system of the formant-tracking type can be considered to have practical value.

Among the topics for research are the following: (a) Procedures for extraction of formant frequencies during non-nasal voiced sounds require further study. Part of this study involves the establishment of suitable criteria for the accuracy of formant tracking, particularly during rapid changes in formant frequencies at vowel boundaries. (b) Methods for synthesizing and for extracting appropriate parametric representations of speech sounds that are generated by vocal-tract excitation at points other than the glottis or that are characterized by nasal consonants need to be examined in detail.

Inability of the analyzer to track rapid formant transitions and to provide a proper specification of consonant spectra seems to be a basic limitation of present formant-tracking systems. Superimposed upon these errors are deteriorations in intelligibility and voice quality arising from inadequate reproduction of the voice excitation. As noted previously , this difficulty is encountered in any system that requires tracking of the fundamental frequency.

As a result of concentrated research effort it may eventually be possible to realize a high-performance formant vocoder. It may happen, however, that high performance can only be achieved at the expense of more complex equipment, possibly requiring some delay in order to perform the required operations on the speech signal.

Peak-picker.[44] The peak-picker may be considered to be a modification of the conventional channel vocoder, although it has some features of the formant vocoder. At every instant of time special peak-detection circuitry determines those few filters in a filter bank which have the greatest outputs. Signals specifying these filters and the magnitudes of their outputs, together with the usual source information, are transmitted to the receiver. At the synthesizer only those modulators which correspond to the selected filters are in operation, and thus peakpicking allows an additional reduction in bandwidth over the conventional channel vocoder. The intelligibility and voice quality of speech processed in this manner are somewhat inferior when compared to the test results obtained for a well-engineered channel vocoder. It is probable, therefore, that the future potential of the peak-picker, at least in its original form, is not great.

4.6 Group E: Below 800 bits/sec (Under 100 cps)

Spectrum Pattern Matching and Coding. C. P. Smith has described a technique whereby the channel capacity required for the channel vocoder can be reduced through a spectrum matching and coding process.[48,49] This technique involves the quantization of incoming signals in frequency and time, and the coding of successive spectral samples in terms of a limited catalog of stored speech patterns. This pattern-matching technique requires an estimated information rate of only 400 to 800 bits/sec. Although the technique has not yet been tested, it is expected to give a speech output comparable to that of the conventional digitized channel vocoder. A disadvantage of the method is that a large, rapid-access memory is necessary; hence application is restricted to communication links with elaborate terminal facilities. In view of the substantial compression achieved, however, it is clear that this technique has considerable potential. Further research and equipment development are being carried out by the Air Force Cambridge Research Laboratories and their contractors.

4.7 Current Research Relevant to the Development of Speech Compression Systems

A number of research studies that are currently in progress are helping to contribute to our knowledge of the speech communication process. It is appropriate in this report to speculate on the potential application of this research to the future development of practical speech compression systems. It is suggested that speech compression systems that are characterized by low information rates (less than, say, 2000 bits/sec) cannot be developed to the point where the speech quality approaches that of conventional systems until some of these research studies have yielded a better understanding of the human speech process. Furthermore, the research should lead to techniques for improving the performance of compression systems with higher information rates, say in the range 2000 - 5000 bits /sec.

In the following paragraphs brief descriptions of some of the currently active research projects are given. The topics include studies of the nature of human generation and perception of speech and studies of new methods of speech analysis and synthesis.

Inverse filtering. One research item that is relevant to formant vocoder systems is the study of a speech analysis procedure known as inverse filtering.[28,34,39] The procedure can be considered to be an application of a general analysis technique that has been called "analysis by synthesis."[19] The method can be used to obtain rather precise measures of the frequency positions of the poles and zeros of the vocal-tract transfer function during vowel and consonant utterances. Procedures for extracting parameters describing these pole and zero locations as a function of time have an important bearing on the development of speech compression systems, since such parameters are known to provide a compact description of the speech signal. Basically the inverse filtering procedure requires that a set of filters be adjusted automatically in such a way that the transfer function of the filters is the reciprocal of the transfer function of the vocal tract that was used to generate the signal. The source of vocal-tract excitation then appears at the output of the inverse filter. An alternative procedure is to perform the operations in the frequency domain by finding a set of pole and zero locations that yield a spectrum that matches the speech spectrum under analysis.[2]

These inverse filtering methods have not yet been realized in a real-time situation. If real-time analysis of this type can be achieved, then its incorporation into the analyzer of a formant vocoder should greatly improve the performance of the system. The analysis procedure is likely to require rather complex operations, however, and some delay or memory will probably be required.

Nature of glottal wave. It has been reasonably well established that the voice quality of both synthetic and natural speech is determined largely by the nature of the glottal excitation. In order to obtain good quality for synthetic speech, the waveform of each glottal excitation pulse must have the proper shape, and the proper temporal relations must exist between successive glottal pulses during an utterance.[3,14] Several studies of these and other aspects of glottal excitation are in progress, and the results of these studies should lead to suggestions for the improvement of voice quality of vocoder systems, including specifications for the design of improved pitch extractors.[16,17,25,26]

In one group of studies, a careful examination is being made of the waveform of the volume velocity output of the glottis, using inverse filtering techniques.[3,34] These studies show that the waveform has a triangular appearance, and hence the spectrum of the glottal output is characterized by a set of zeros located close to the $j\omega$-axis in the complex frequency plane. A more fundamental approach is being followed by groups that are examining in detail the mechanism of operation of the larynx through photographic and other techniques.[50,55] These studies also show that the area of the glottis opening has a triangular waveshape, but that this waveform varies with voice effort and with fundamental frequency.

Measurements of the intervals between successive glottal pulses have shown that a certain amount of quasi-random pulse position modulation is superimposed on the smooth changes in inter-pulse interval associated with varying inflection patterns.[32] Perceptual experiments with synthetic speech have demonstrated that the voice quality is improved if such randomnesses are superimposed on the regular inflection patterns.[25,26]

Pitch extraction. Although the first pitch extractor was devised several decades ago, effort continues to be devoted to the development of a device that indicates the position of each glottal pulse with a minimum number of errors. The basic problem is to devise a procedure that operates in a satisfactory manner for a wide range of fundamental frequencies and for many different speakers and different voice efforts. Most of the schemes that have been under study recently have attempted to find the location of each glottal pulse by making a number of separate determinations of the pulse location through a series of measurements on the waveform or on some transformed version of the waveform, and then making a decision on the presence or absence of a pulse by looking for a coincidence among several of the separate determinations.[16,17,18,45]

It is evident that research on methods for pitch extraction should go hand in hand with basic studies of the glottal wave in natural speech and with studies of human pitch perception. The psychoacoustic studies should help to establish criteria that indicate whether a pitch extractor is operating in a satisfactory manner.

Articulatory synthesizer. As discussed above, a limitation of the present formant vocoder technique stems from the fact that the synthesizer is designed primarily for the generation of vowels and vowel-like sounds, and it is necessary to generate certain consonant sounds by means of special circuits or through some other tour de force. Some effort is currently being devoted to the development of a synthesizer that is in principle capable of generating all vowels and consonants with a single time-varying network. This synthesizer is an analog of the acoustic tube that forms the vocal tract between the glottis and the lips, including the nasal cavities.[10,21,46,52] Changes in vocal-tract configuration are simulated in the synthesizer through control of the values of a set of variable electrical elements. The analog circuit can be excited by an electrical buzz source at one end, simulating glottal excitation, or by an electrical noise source at some point along its length, simulating excitation by noise that results from turbulence in the vocal tract. Thus voiced and voiceless sounds are generated by the same circuit simply by changing the location and spectrum of the source.

Present research is devoted to finding the nature of the control signals that must be applied to such a synthesizer in order to generate natural speech. Such research must necessarily include studies of the actual vocal-tract configurations used in natural speech and the motions of the anatomical components that give rise to these configurations.[40,54] In order to minimize the information rates of the signals controlling the synthesizer, means must be found for describing the articulatory activities with a relatively small number of parameters.[53] It is probable that the information rates associated with the control signals for such a synthesizer would be equal to or less than those needed to control the synthesizer in a formant vocoder, i.e., 1000 bits/sec or less.

Articulatory analysis. If a speech compression system using an articulatory synthesizer is contemplated, it will be necessary at the transmitting end to extract from the speech wave a set or control signals that describe in some approximate manner the configurations and excitations of the vocal tract. Research toward these objectives has hardly begun, but it is evident that a reasonably complex set of calculations will have to be made in order to extract the proper signals.[22]

Acoustic properties of speech sounds. While it is known that certain classes of speech sounds, particularly the non-nasal vowels, can be described rather precisely yet compactly in terms of the frequencies of the first two or three formants, comparable procedures for the description of other classes of sounds are still under study. Such procedures are relevant to the development of compression systems of the formant vocoder type, since it is desirable to devise methods for properly generating various classes of sounds at the receiver and for extracting from the input speech signal a set of parameters that can be used to control the synthesizer.

Studies have shown, for example, that the spectra of fricative consonants can be described approximately by one or two poles and one zero,[23] but methods for automatically extracting these parameters from the speech signal have not yet been devised. Similarly, the spectra of nasal consonants are characterized by a number of poles and zeros, although the synthesis of consonants in this class can be approximated by using a circuit whose transfer function has only poles if the bandwidths of the resonances are made sufficiently broad.[41] Further studies are necessary, however, before these and other classes of consonants are understood sufficiently well that they can be handled properly in a formant vocoder system.

Cues for identification of speech sounds. Over a period of years a number of studies have been carried out to determine which features of the acoustic speech signal constitute the principal cues used by a listener to identify the signal. For example, experiments have led to an approximate specification of the directions and rates of change of the formant transitions between vowels and stop and nasal consonants.[29,30] These studies have particular significance for the design of speech compression systems, since they indicate the types of distortions that are likely to obscure certain cues and thus lead to loss of intelligibility.[4] Thus, as the results of these perceptual studies become available, they should suggest how present speech compression systems can be improved and they will provide a basis for the design of future systems.

5. APPENDICES

5.1 Appendix I - References

1. A.S.A. S3.2-1960. American standard measurement for monosyllabic word intelligibility.

2. Bell, C. G., Fujisaki, H., Heinz, J. M., Stevens, K. N., and House, A. S., Reduction of speech spectra by analysis-by-synthesis techniques. J. Acoust. Soc. Am. 33, 1725-1736 (1961).

3. Cederlund, C., Krokstad, A., and Kringlebotn, M., Voice source studies. Quart. Prog. Status Report, Speech Transmission Laboratory, Royal Inst. of Technol. (Stockh.), Oct. 1960; also, Air Force Cambridge Research Lab. Document AFCRL-390, Jan. 1961.

4. Cooper, F. S., Basic factors in speech perception and applications to speech processing. Proc. Sem. Speech Compression and Processing, Air Force Cambridge Research Center TR-59-198, Sept. 1959, Vol. I.

5. Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J., Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Am. 24, 597-606 (1952).

6. David,Jr., E. E., Naturalness and distortion in speech-processing devices. J. Acoust. Soc. Am. 28, 586-589 (1956).

7. Delattre, P. C., Liberman, A. M., and Cooper, F. S., Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. 27, 769-773 (1955).

8. Dudley, H. W., The Vocoder. Bell Lab. Rec. 18, 122-126 (1936).

9. Dudley, H. W., The carrier nature of speech. Bell Syst. Tech. J. 19, 495-515 (1940).

10. Dunn, H. K., The calculation of vowel resonances, and an electrical vocal tract. J. Acoust. Soc. Am. 22, 740-753 (1950).

11. Egan, J. P., Articulation testing methods. Laryngoscope 58, 955-961 (1948).

12. Fant, C. G. M., Acoustic Theory of Speech Production (Mouton Co., 's-Gravenhage, 1960).

13. Fant, C. G. M., and Stevens, K. N., Systems for speech compression. Fortschritte der Hochfrequenztechnik, 5, Akademische Verlagsgesellschaft m. b. H., Frankfurt/a.M., 229-262 (1960).

14. Flanagan, J. L., Some properties of the glottal sound source. J. Speech Hearing Res., 1, 99-116 (1958).

15. Fujimura, O., Some synthesis experiments on stop consonants in the initial position. Quart. Prog. Rept. 61, Research Laboratory of Electronics, Mass. Inst. of Technol., April 1961, pp. 153-162.

16. Gill, J. S., Automatic extraction of the excitation function of speech with particular reference to the use of correlation methods. Proc. 3rd Int. Congr. Acoustics, ed. L. Cremer, Elsevier Publ. Co., New York, 1961, Vol. I, pp. 217-220.

17. Gill, J. S., A study of the requirements for excitation control in synthetic speech. Proc. 3rd Int. Congr. Acoustics, ed. L. Cremer, Elsevier Publ. Co., New York, 1961, Vol. I, pp. 221-224.

18. Gold, B., Pitch extraction on the TX-2 computer. J. Acoust. Soc. Am., 33, 1664-1665 (A) (1961).

19. Halle, M., and Stevens, K. N., Analysis by synthesis. Proc. Sem. Speech Compression and Processing, Air Force Cambridge Research Center TR-59-198, Sept. 1959, Vol. II.

20. Harris, K. S., Cues for the identification of the fricatives of American English. Language and Speech 1, 1 (1958).

21. Hecker, M. H. L., Studies of nasal consonants with an articulatory speech synthesizer. J. Acoust. Soc. Am. 34, 179 (1962).

22. Heinz, J. M., Reduction of speech spectra to descriptions in terms of vocal-tract area functions. Quart. Prog. Rept. 64, Research Laboratory of Electronics, Mass. Inst. of Technol., Jan. 1962, pp. 198-203.

23. Heinz, J. M., and Stevens, K. N., On the properties of voiceless fricative consonants. J. Acoust. Soc. Am. 33, 589-596 (1961).

24. House, A. S., Stevens, K. N., and Fujisaki, H., Automatic measurement of the formants of vowels in diverse consonatal environments. J. Acoust. Soc. Am. 32, 1517 (A) (1960).

25. Kersta, L. G., Bricker, P. D., and David, Jr., E. E., Human or machine? -- A study of voice naturalness. J. Acoust. Soc. Am. 32, 1502 (A) (1960).

26. Kleinschmidt, K., The effect of quasi-periodicity on the naturalness of synthetic vowels. S.B. thesis, Dept. Elec. Eng., M.I.T., 1957.

27. Kryter, K. D., Speech bandwidth compression through spectrum selection. J. Acoust. Soc. Am. 32, 547-556 (1960).

28. Lawrence, W., Formant tracking by self-adjusting inverse filtering. J. Acoust. Soc. Am. 33, 1676 (A) (1961).

29. Liberman, A. M., Some results of research on speech perception. J. Acoust. Soc. Am. 29, 117-123 (1957).

30. Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J., The role of consonant-vowel transitions in the perception of stop and nasal consonants. Psychol. Monogr. 68, No. 8, 1-13 (1954).

31. Licklider, J. C. R., The intelligibility of amplitude-dichotomized time-quantized speech waves. J. Acoust. Soc. Am. 22, 820-823 (1950).

32. Lieberman, P., Perturbations in vocal pitch. J. Acoust. Soc. Am. 33, 597-603 (1961).

33. Mathews, M. V., Extremal coding for speech transmission. Trans. Inst. Radio Engrs. IT-5, 129-136 (1959).

34. Mathews, M. V., Miller, J. E., and David, Jr., E. E., Pitch synchronous analysis of voiced sounds. J. Acoust. Soc. Am. 33, 179-186 (1961).

35. Marcou, P., and Daguet, J., New methods of speech transmission. Information Theory, ed. C. Cherry, Butterworths Scientific Publications, London 1956, pp. 231-244.

36. Marill, T., Automatic recognition of speech. Rome Air Development Center TN-60-196, Oct. 1960.

37. Miller, G. A., Heise, G. A., and Lichten, W., The intelligibility of speech as a function of the context of the test materials. J. Exp. Psychol. 41, 329-335 (1951).

38. Miller, G. A., and Nicely, P. E., Analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338-352 (1955).

39. Miller, R. L., Nature of the vocal cord wave. J. Acoust. Soc. Am. 31, 667-677 (1959).

40. Moll, K. L., Cinefluorographic techniques in speech research, J. Speech Hearing Res. 30, 227-241 (1960).

41. Nakata, K., Synthesis and perception of nasal consonants. J. Acoust. Soc. Am. 31, 661-666 (1959).

42. O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., and Cooper, F. S., Acoustic cues for the perception of initial /w, j, r, ℓ/ in English. Word 13, No. 1, 24-43, (1957).

43. Peterson, G. E., and Barney, H. L., Control methods used in a study of the vowels. J. Acoust. Soc. Am. 24, 175-184 (1952).

44. Peterson, E., and Cooper, F. S., Peakpicker: a band-width compression device. J. Acoust. Soc. Am. 29, 777 (A) (1957).

45. Risberg, A., Voice fundamental frequency tracking. Quart. Prog. Status Report, Speech Transmission Laboratory, Royal Inst. of Technol. (Stockh.), April 1961.

46. Rosen, G., A dynamic analog speech synthesizer. J. Acoust. Soc. Am. 30, 201-209 (1958).

47. Schroeder, M. R., and David, Jr., E. E., A vocoder for transmitting 10 kcps speech over a 3.5 kcps channel. Acustica 10, 35-43 (1960).

48. Smith, C. P., Speech data reduction: voice communications by means of binary signals at rates under 1000 bits per second. Air Force Cambridge Research Center TR-57-111, Jan. 1957; also ASTIA Document No. AD 117290.

49. Smith, C. P., A method for speech data processing by means of a digital computer. Air Force Cambridge Research Center ERD-TM-58-103, 1958.

50. Sonesson, B., On the anatomy and vibratory pattern of the human vocal folds. Acta Otolaryngologica, Suppl. 156, 7-80 (1960).

51. Stevens, K. N., Review of existing speech compression systems. Rome Air Development Center TN-60-197, Oct. 1960.

52. Stevens, K. N., Kasowski, S., and Fant, G., An electrical analog of the vocal tract. J. Acoust. Soc. Am. 25, 734-742 (1953).

53. Stevens, K. N., and House, A. S., Development of a quantitative description of vowel articulation. J. Acoust. Soc. Am. 27, 484-493 (1955).

54. Truby, H. M., Acoustic-cineradiographic analysis considerations with special reference to certain consonantal complexes. Acta Radiol., Suppl. 182 (1959).

55. van den Berg, J., and Tan, T. S., Results of experiments with human larynxes. Practica Oto-Rhino-Laryngol., 21, 425-450 (1959).

5.2 Appendix II - Statistical Analyses

The experimental procedures used in the PB word intelligibility and talker identification tests were arranged to permit the determination of some overall measure of significance in the form of an analysis of variance. In the case of the intelligibility data a three-way analysis scheme was used. The arrangement may be visualized in the form of a cube with the systems under test on one axis, the talkers who read the test materials on another axis, and the listeners (subjects) on the third axis. The talker identification tests used a two-way analysis scheme in which the systems under test were on one axis and the listeners on another.

Table 5.2-1 shows the analysis of variance of the intelligibility scores obtained for seven systems, using two male talkers and eight listeners. When the triple interaction variance term is used as the demoninator in an F ratio, none of the simple interactions reach a magnitude that is significant at the 1% level of confidence. Similar tests of the main effects demonstrate that the variance attributable to Systems and to Listeners is statistically significant ($p \leq .01$); there is no marked effect of Talkers in this study.

Table 5.2-1
Analysis of Variance of Intelligibility Scores
Obtained for Seven Systems Using Two Talkers
and Eight Listeners

Source of Variation	Sum of Squares	d.f.	Variance	F*
Systems (S)	41,305.11	6	6,884.19	330.34**
Talkers (T)	89.26	1	89.26	4.28
Listeners (L)	539.99	7	77.14	3.70**
T x S	340.55	6	56.76	2.72
T x L	139.31	7	19.90	-
S x L	1,295.82	42	30.85	1.48
T x S x L	875.38	42	20.84	-
Total	44,585.42	111		

* $\underline{F} = V/V_{TxSxL}$

** $p \leq 0.01$ level of confidence

The finding that a significant variance is attributable to Systems permits an examination of the differences among the system scores. The results of t tests in Table 5.2-2 demonstrate that almost all of the ranked system scores differ significantly from contiguous scores. (It is probable that the Stromberg system differs significantly from the Tasaroff-Daguet system, at least at the p = 0.05 level of confidence.)

Table 5.2-2

Mean Intelligibility Scores and *t* Tests
Arranged to Demonstrate Significant Differences Among Systems
(Seven Systems, Two Talkers and Eight Listeners)

Systems	M Score	M Diff.	*t*
Reference	95		
		9	8.46*
Stromberg	86		
		1	1.25
Philco	85		
		6	3.20*
Tasaroff-Daguet	79		
		11	6.19*
Narrow Band	68		
		7	4.93*
Hughes	61		
		28	13.33*
Melpar	33		

* $p \leq 0.01$ level of confidence

A similar kind of analysis was made with five systems using four talkers (two male and two female) and eight listeners. The results of this analysis are shown in Table 5.2-3. In this analysis the Talker x Systems interaction was significantly large and was used to test certain main effects. This interaction indicates that the score obtained from a given system was affected significantly by differences among the talkers. The major sources of variation, however, were contributed by the Systems, the Talkers, and the Listeners (main effects).

Table 5.2-3
Analysis of Variance of Intelligibility Scores
Obtained for Five Systems Using Four Talkers
and Eight Listeners

Source of Variation	Sum of Squares	d.f.	Variance	F*
Systems (S)	30,422.75	4	7,605.68	33.62**
Talkers (T)	22,227.47	3	7,409.15	32.75**
Listeners (L)	1,225.32	7	175.04	13.62**
T x S	2,714.90	12	226.24	17.61**
T x L	456.18	21	21.72	1.69
S x L	622.30	28	22.22	1.73
T x S x L	1,079.45	84	12.85	-
Total	58,748.37	159		

* $\underline{F} = V_{TxS}/V_{TxSxL}$; V_{TxL}/V_{TxSxL}; V_{SxL}/V_{TxSxL}; V_{L}/V_{TxSxL}; V_{S}/V_{TxS}; V_{T}/V_{TxS}

** $p \leq 0.01$ level of confidence

The significant differences among the five systems are specified in Table 5.2-4, which supports the findings of the earlier analysis (Table 5.2-2).

Table 5.2-4
Mean Intelligibility Scores and *t* Tests Arranged to Demonstrate Significant Differences Among Systems (Five Systems, Four Talkers and Eight Listeners.)

Systems	M score	M diff	*t*
Reference	89		
		17	5.61*
Philco	72		
		1	1.19
Stromberg	71		
		11	10.14*
Narrow Band	60		
		13	9.20*
Hughes	47		

* $p \leq 0.01$ level of confidence

Similar analyses have been accomplished for the talker identification data obtained with Quartets I and II. In this case, however, the scores were arranged to test the contributions to the total variance of the Systems and the Listeners. The results of these two-way analyses of variance are presented in Tables 5.2-5 and 5.2-7. For both quartets there were statistically significant differences among the talker identification scores obtained from the systems under test, but the ranges of scores were smaller than those obtained in the intelligibility tests. The mean talker identification scores obtained for Quartets I and II, along with the results of *t* tests between adjacent scores, are shown in Tables 5.2-6 and 5.2-8, respectively.

Table 5.2-5

Analysis of Variance of Talker Identification Scores Obtained with Quartet I, using Seven Systems and 29 Listeners

Source of Variation	Sum of Squares	d.f.	Variance	F*
Systems	551.77	6	91.96	17.96**
Listeners	462.40	28	16.51	3.23
Remainder	859.94	168	5.12	--
Total	1,874.11	202		

*$F = V/V_{Rem.}$

** $p \leq 0.01$ level of confidence

Table 5.2-6

Mean Talker Identification Scores and t Tests for Quartet I Arranged to Demonstrate Significant Differences Among Systems (Seven Systems and 29 Listeners.)

Systems	M score	M diff	t
Reference	54		
		9	2.69*
Tasaroff-Daguet	45		
		0	--
Narrow Band	45		
		2	--
Philco	43		
		2	--
Stromberg	41		
		8	3.23*
Hughes	33		
		6	1.96*
Melpar	27		

* $p \leq 0.05$ level of confidenc

Table 5.2-7

Analysis of Variance of Talker Identification Scores Obtained with Quartet II, using Seven Systems and 30 Listeners

Source of Variation	Sum of Squares	d.f.	Variance	F*
Systems	731.63	6	121.94	19.05**
Listeners	505.42	29	17.43	2.72
Remainder	1,113.20	174	6.40	--
Total	2,350.25	209		

*F = $V/V_{Rem.}$

** $p \leq 0.01$ level of confidence

Table 5.2-8

Mean Talker Identification Scores and t Tests for Quartet II Arranged to Demonstrate Significant Differences Among Systems (Seven Systems and 30 Listeners.)

System	M score	M diff	t
Reference	59		
		5	1.22
Stromberg	54		
		6	2.33*
Narrow Band	48		
		4	1.46
Tasaroff-Daguet	44		
		7	2.69*
Melpar	37		
		2	--
Philco	35		
		4	--
Hughes	31		

* $p \leq 0.05$ level of confidence

Rome Air Development Center, Griffiss AF Base, New York
Report No. RADC-TDR-62-171 AN EVALUATION OF SPEECH COMPRESSION SYSTEMS. Interim report, 1 Mar 62, 129 p. incl tables.

Unclassified Report

The results of PB word and nonsense syllable intelligibility tests, voice quality, talker identification, and continuous speech tests of selected speech compression systems are presented. The systems were: a "reference" low-pass (approx. 3000 cps) filter system, two channel vocoders, a semi-vocoder, a formant-tracking vocoder and a multiple narrow band filter system. The status of various speech compression techniques, current relevant research and recommendations for future research and development in this area are reported. Different speech compression techniques are classified according to their ability to provide a given level of speech intelligibility at different information rates.

It is judged that channel vocoders operating at about 2400 bits/sec and semi-vocoders at an estimated 9600 bits/sec provide adequate intelligibility and quality for most military communications; the quality of the semi-vocoder is superior to the channel vocoder. Formant-tracking vocoders utilize the lowest information rate (about 1000 bits/sec) of any of the bandwidth compression techniques tested. Formant-tracking vocoders require further improvement before they can be considered as satisfactory for general use.

1. Speech transmission
2. Systems for bandwidth compression of speech
3. Channel vocoder
4. Semi-vocoder
5. Formant-tracking vododer
6. Narrow-band speech system

I. Project No. 4519
Task No. 45350
II. Contract AF 30(602)-2235
III. Bolt Beranek and Newman Inc.
Cambridge, Mass.
IV. Stevens, K. N., M. H. Hecker,
K. D. Kryter
V. BBN Report No. 914
VI. TDR-62-171
VII. In ASTIA collection

Rome Air Development Center, Griffiss AF Base, New York
Report No. RADC-TDR-62-171 AN EVALUATION OF SPEECH COMPRESSION SYSTEMS. Interim report, 1 Mar 62, 129 p. incl tables.

Unclassified Report

The results of PB word and nonsense syllable intelligibility tests, voice quality, talker identification, and continuous speech tests of selected speech compression systems are presented. The systems were: a "reference" low-pass (approx. 3000 cps) filter system, two channel vocoders, a semi-vocoder, a formant-tracking vocoder and a multiple narrow band filter system. The status of various speech compression techniques, current relevant research and recommendations for future research and development in this area are reported. Different speech compression techniques are classified according to their ability to provide a given level of speech intelligibility at different information rates.

It is judged that channel vocoders operating at about 2400 bits/sec and semi-vocoders at an estimated 9600 bits/sec provide adequate intelligibility and quality for most military communications; the quality of the semi-vocoder is superior to the channel vocoder. Formant-tracking vocoders utilize the lowest information rate (about 1000 bits/sec) of any of the bandwidth compression techniques tested. Formant-tracking vocoders require further improvement before they can be considered as satisfactory for general use.

1. Speech transmission
2. Systems for bandwidth compression of speech
3. Channel vocoder
4. Semi-vocoder
5. Formant-tracking vododer
6. Narrow-band speech system

I. Project No. 4519
Task No. 45350
II. Contract AF 30(602)-2235
III. Bolt Beranek and Newman Inc.
Cambridge, Mass.
IV. Stevens, K. N., M. H. Hecker,
K. D. Kryter
V. BBN Report No. 914
VI. TDR-62-171
VII. In ASTIA collection

Rome Air Development Center, Griffiss AF Base, New York
Report No. RADC-TDR-62-171 AN EVALUATION OF SPEECH COMPRESSION SYSTEMS. Interim report, 1 Mar 62, 129 p. incl tables.

Unclassified Report

The results of PB word and nonsense syllable intelligibility tests, voice quality, talker identification, and continuous speech tests of selected speech compression systems are presented. The systems were: a "reference" low-pass (approx. 3000 cps) filter system, two channel vocoders, a semi-vocoder, a formant-tracking vocoder and a multiple narrow band filter system. The status of various speech compression techniques, current relevant research and recommendations for future research and development in this area are reported. Different speech compression techniques are classified according to their ability to provide a given level of speech intelligibility at different information rates.

It is judged that channel vocoders operating at about 2400 bits/sec and semi-vocoders at an estimated 9600 bits/sec provide adequate intelligibility and quality for most military communications; the quality of the semi-vocoder is superior to the channel vocoder. Formant-tracking vocoders utilize the lowest information rate (about 1000 bits/sec) of any of the bandwidth compression techniques tested. Formant-tracking vocoders require further improvement before they can be considered as satisfactory for general use.

1. Speech transmission
2. Systems for bandwidth compression of speech
3. Channel vocoder
4. Semi-vocoder
5. Formant-tracking vododer
6. Narrow-band speech system

I. Project No. 4519
Task No. 45350
II. Contract AF 30(602)-2235
III. Bolt Beranek and Newman Inc.
Cambridge, Mass.
IV. Stevens, K. N., M. H. Hecker,
K. D. Kryter
V. BBN Report No. 914
VI. TDR-62-171
VII. In ASTIA collection

Rome Air Development Center, Griffiss AF Base, New York
Report No. RADC-TDR-62-171 AN EVALUATION OF SPEECH COMPRESSION SYSTEMS. Interim report, 1 Mar 62, 129 p. incl tables.

Unclassified Report

The results of PB word and nonsense syllable intelligibility tests, voice quality, talker identification, and continuous speech tests of selected speech compression systems are presented. The systems were: a "reference" low-pass (approx. 3000 cps) filter system, two channel vocoders, a semi-vocoder, a formant-tracking vocoder and a multiple narrow band filter system. The status of various speech compression techniques, current relevant research and recommendations for future research and development in this area are reported. Different speech compression techniques are classified according to their ability to provide a given level of speech intelligibility at different information rates.

It is judged that channel vocoders operating at about 2400 bits/sec and semi-vocoders at an estimated 9600 bits/sec provide adequate intelligibility and quality for most military communications; the quality of the semi-vocoder is superior to the channel vocoder. Formant-tracking vocoders utilize the lowest information rate (about 1000 bits/sec) of any of the bandwidth compression techniques tested. Formant-tracking vocoders require further improvement before they can be considered as satisfactory for general use.

1. Speech transmission
2. Systems for bandwidth compression of speech
3. Channel vocoder
4. Semi-vocoder
5. Formant-tracking vododer
6. Narrow-band speech system

I. Project No. 4519
Task No. 45350
II. Contract AF 30(602)-2235
III. Bolt Beranek and Newman Inc.
Cambridge, Mass.
IV. Stevens, K. N., M. H. Hecker,
K. D. Kryter
V. BBN Report No. 914
VI. TDR-62-171
VII. In ASTIA collection